Universidad Autónoma de Madrid
Escuela politécnica superior

Máster en Bioinformática y Biología Computacional

Trabajo Fin de Máster

APLICACIÓN DE TÉCNICAS DE NLP Y MAPEO ONTOLÓGICO EN EL ÁMBITO DEL DIAGNÓSTICO CLÍNICO

Autor: Juan Antonio Miguel González
Tutor: Paolo Maietta
Ponente: Daniel Hernández Lobato

Junio 2019

APLICACIÓN DE TÉCNICAS DE NLP Y MAPEO ONTOLÓGICO EN EL ÁMBITO DEL DIAGNÓSTICO CLÍNICO

Autor: Juan Antonio Miguel González
Tutor: Paolo Maietta
Ponente: Daniel Hernández Lobato

NIMGenetics
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Junio 2019

Resumen

Resumen

En la actualidad la rutina de trabajo de los servicios hospitalarios incluye el uso de técnicas de secuenciación masiva para el análisis y estudio de las posibles causas genéticas que provocan la enfermedad del paciente. En el ámbito del diagnóstico de enfermedades raras, la implantación de técnicas de secuenciación masiva como *Whole Exome Sequencing* (WES) han mejorado la rentabilidad y reducido el tiempo de diagnóstico en este tipo de desórdenes. Este análisis genera un listado de variantes a partir del cual el analista selecciona aquella variante con mayor probabilidad de causar el fenotipo del paciente. Para agilizar este proceso, la historia clínica del paciente debería establecer de forma precisa y estandarizada la descripción del fenotipo, para así poder realizar un análisis más dirigido respecto de la enfermedad que presente. Uno de los problemas que dificulta este objetivo radica en que los informes clínicos suelen estar escritos en lenguaje natural humano, lo que dificulta su implementación algorítmica. En este contexto, la *Human Phenotype Ontology* (HPO) proporciona una terminología médica estandarizada en relación con las anomalías fenotípicas humanas, la cual puede combinarse con modelos de inferencia como exomiser. Este algoritmo es capaz de priorizar una lista de variantes utilizando el VCF generado tras la secuenciación y los términos HPO. En esta lista cada variante lleva asociada una puntuación, la cual es adjudicada en función de la posible causalidad de la misma respecto del fenotipo descrito. En este trabajo se diseñó un pipeline completo para la extracción de los conceptos médicos incluidos en los informes de clínica, los cuales fueron asociados posteriormente con su terminología HPO correspondiente de forma manual o mediante mapeo automático. La ejecución de exomiser permitió establecer una comparativa entre la lista de variantes priorizada utilizando los términos asociados manualmente y los generados mediante mapeo automático. Además, también se testeó y comprobó la funcionalidad de la herramienta *Health29* desarrollada por la *Fundacion29*. Los resultados obtenidos permitieron afirmar que los procedimientos desarrollados podrían ser útiles como soporte analítico para la identificación y validación de variantes genómicas en el ámbito del diagnóstico clínico.

Palabras Clave

Next Generation Sequencing, Whole Exome Sequencing, Human Phenotype Ontology, Natural Language Processing, Text Mining, Exomiser.

Abstract

Currently hospital services include high throughput sequencing for the analysis and the study of genetic background in patient disease. Next generation sequencing techniques have made it possible by improving the efficiency and reducing time needed to diagnose rare diseases. WES analysis produces a series of genetic variants with different typology that can be the cause of disease or syndrome, where analyst have to identify the one that really produces the patient phenotype. To make the process faster, patient medical history must contain a precise and standardized description of patient phenotype. Clinical reports use to be written in human natural language, what makes more difficult the algorithm performance. In this context, *Human Phenotype Ontology* (HPO) provides a standardized vocabulary of phenotypic abnormalities encountered in human disease, which can be combined with inference models as exomiser. This algorithm is able to make a variant list prioritization using VCF file and HPO terms. Each genetic variant included in this list is associated to a score that exomiser attaches according to its probability of being the patient phenotype cause. In this work it was designed a complete pipeline for extracting medical concepts from clinical reports with their subsequently association to the corresponding HPO terms, by manually or automatic mapping. Exomiser performance allowed to make a comparative between the variant list prioritized by manual association and automatic mapping of HPO terms, including the testing of *Health29* functionality. The results obtained allowed to confirm that developed procedures can be used as an analytic support for genetic variant identification and validation in the field of clinical diagnosis.

Key words

Next Generation Sequencing, Whole Exome Sequencing, Human Phenotype Ontology, Natural Language Processing, Text Mining, Exomiser.

Agradecimientos

Mis agradecimientos van dirigidos principalmente a Paolo Maietta y Marta Carcajona por su gran ayuda y consejos necesarios para guiarme a lo largo de este trabajo. Además de por darme la oportunidad de afrontar un proyecto de estas características, el cual he disfrutado gratamente y me ha permitido aprender mucho respecto de esta disciplina tan exigente y desafiante como es la Bioinformática.

Gracias también a todas mis compañeras y compañeros de NIMGenetics por su amabilidad y por fomentar el buen ambiente de trabajo que se respira en la empresa. Sin olvidar el apoyo recibido por las personas más cercanas.

Reitero una vez más mi agradecimiento por haber podido aportar mi grano de arena en este apasionante proyecto, Gracias.

Índice general

Índice de Figuras

Índice de Tablas

1 Introducción

1.1. Motivación del proyecto

En los últimos años, la práctica clínica y los servicios hospitalarios han ido introduciendo de forma rutinaria las técnicas de secuenciación masiva (NGS) para el estudio del genoma humano. Estas técnicas permiten analizar en detalle el *background* genético de los pacientes y detectar los cambios potenciales que puedan ser causantes de la enfermedad congénita [1]. Uno de los ámbitos médicos que más se ha visto beneficiado respecto de este acontecimiento es el diagnóstico de enfermedades raras.

Las enfermedades raras son aquellas que afectan a un número limitado de individuos (uno de cada 2000 en la Unión Europea y uno cada 1250 en Estados Unidos). Sin embargo, el número estimado de enfermedades de este tipo asciende a más de 5000, donde el verdadero problema reside en la dificultad de su correcto diagnóstico y la reducida disponibilidad de datos epidemiológicos a nivel global [2]. La mayoría de estas enfermedades son trastornos genéticos que afectan sustancialmente a la esperanza y calidad de vida del paciente, por el deterioro de sus capacidades físicas y mentales. La introducción de la NGS ha supuesto un aumento de la rentabilidad y reducción de los tiempos de diagnóstico para este tipo de desórdenes, los cuales serían muy difíciles de confirmar únicamente por medios estrictamente clínicos, como por ejemplo mediante la búsqueda de patrones sintomáticos compartidos entre distintos pacientes [3].

Los pasos característicos que integra un *workflow* de NGS para el análisis de pacientes con enfermedades raras son los siguientes: análisis y preprocesamiento del *raw data* obtenido tras la secuenciación, alineamiento de secuencias (mapeo de lecturas respecto de un genoma de referencia), post-procesamiento del mapeo para minimizar los posibles artefactos generados durante el alineamiento y, por último, el análisis de variantes, que incluye el *variant calling* y la anotación y priorización de variantes [4]. El *variant calling* genera un archivo VCF donde se observan las diferencias en los alineamientos de las lecturas respecto del genoma de referencia, incluyendo cualquier tipo de variación a nivel secuencial (SNPs, indels...) [5]. La evolución en el desarrollo de los algoritmos utilizados en el *variant calling* ha incrementado el rendimiento y la sensibilidad del proceso, además de la reducción de la tasa de falsos positivos y falsos negativos que pueden introducirse en la lista de variantes [3]. El aumento exponencial de datos de secuenciación genéticos ha permitido aumentar también la fiabilidad de los procesos de anotación de las características genéticas en las variantes identificadas [6]. La última etapa del proceso,

la priorización de las variantes, consiste en aplicar una estrategia cuyo objetivo sea construir una lista significativa de variantes candidatas para su validación experimental, filtrando aquellas variantes con menor probabilidad de estar relacionadas con la enfermedad del paciente. Estas variantes susceptibles de ser eliminadas son por ejemplo aquellas que tienen una baja cobertura y calidad en el alineamiento o que han sido previamente reportadas como variante común a nivel poblacional [4][7].

El resultado final tras el análisis varía ampliamente. En las enfermedades mejor caracterizadas la posibilidad de identificar la variante causal será más fácil por el hecho de que existe más información registrada en relación con ese tipo de enfermedad. Sin embargo, estas situaciones son poco usuales, debido a que en la mayoría de los casos, el número de variantes candidatas encontradas en un estudio genético de este tipo es muy elevado, por lo que una caracterización fenotípica bien realizada será crucial para elaborar la lista de posibles variantes candidatas. En esta lista se pueden encontrar variantes genéticas de diversa tipología, siendo la mayoritaria la variante de significado incierto (VUS) y después la de tipo *patogénica*, aunque este hecho no implica necesariamente que esta variante sea la causante del fenotipo del paciente [8]. Es por ello que el apoyo a nivel fenotípico es esencial para elaborar un diagnóstico final y agilizar el proceso de validación [9]. Este hecho se puede corroborar teniendo en cuenta que, aunque existan algoritmos diseñados para caracterizar y proporcionar una puntuación de las variantes encontradas [10], y compañías especializadas en el desarrollo de software profesional, la búsqueda de estas variantes dañinas sigue siendo un proceso muy costoso a nivel de tiempo y que también depende de la experiencia del analista que la esté llevando a cabo.

La historia clínica del paciente se considera información clave para el éxito del análisis porque permite reducir la magnitud del estudio y centrarse en aquellos genes cuya relación con la enfermedad haya sido establecida previamente [1]. En los casos en que la enfermedad se encuentra bien caracterizada, el análisis estará centrado en pocos genes, mientras que si la enfermedad a diagnosticar es más compleja y se dispone de menos información, lo ideal es aproximarse a un enfoque analítico del exoma del paciente. Por ello, siempre es recomendable que los médicos que soliciten un análisis de tipo WES, consideren detallar y describir el fenotipo del paciente con al menos un número de síntomas específicos de entre 3 y 10 [11], para evitar que se produzcan las siguientes situaciones: i) detallar un número insuficiente de síntomas y, como consecuencia, obtener un número de genes candidatos muy elevado, o lo que es lo mismo, obtener un número excesivo de VUSes [12] en estos casos. ii) exceder el número de síntomas necesarios y reducir las probabilidades de los resultados del análisis a una enfermedad monogénica, omitiendo las enfermedades genéticas de etiología más compleja. Con todo ello, para justificar un diagnóstico final a partir de WES, los médicos genetistas y facultativos de laboratorio deben establecer un acuerdo en la interpretación de los resultados obtenidos [13].

Por otro lado, sabiendo que la caracterización fenotípica es importante para dirigir el análisis al estudio de genes relacionados con la enfermedad del paciente, es cierto que esta información es presentada por el médico en forma de texto en lenguaje natural humano, lo cual dificulta mucho la implementación algorítmica. Para dar soporte a esta problemática, proyectos como *la Human Phenotype Ontology* (HPO) ofrecen un vocabulario estandarizado de las anomalías fenotípicas encontradas en patologías humanas [14], el cual se posiciona actualmente en números de 123.724 anotaciones de términos HPO para enfermedades raras y 132.620 para enfermedades de tipo común [15] (**Figura 1.1**).

Este avance ha supuesto las bases para el desarrollo de modelos algorítmicos de inferencia como *Exomiser* [16] o *PhenIX* [17], que son capaces de utilizar esta información para estimar la similitud entre los fenotipos descritos y un fenotipo de interés, además de inferir la probabilidad de que un gen asociado con un determinado set de términos HPO pueda estar involucrado con el cuadro clínico que presenta el paciente. El potencial de este enfoque analítico y sus implicaciones a nivel de diagnóstico, tanto por el aumento de la rentabilidad como por la reducción de

tiempos de análisis, motiva la búsqueda de nuevos métodos computacionales capaces de asociar de forma automática términos HPO con información clínica estándar. Sin embargo, los métodos que ya han sido desarrollados, están limitados al idioma Inglés, por lo que por esta razón se ha decidido evaluar su aplicabilidad (y en su defecto proponer alternativas) en el ámbito del idioma Castellano. En este contexto, la empresa NIMGenetics constituye una plataforma ideal para el desarrollo de esta nueva metodología, por tener a disposición del departamento de Bioinformática datos clínicos de casi 20.000 pacientes en sus archivos.

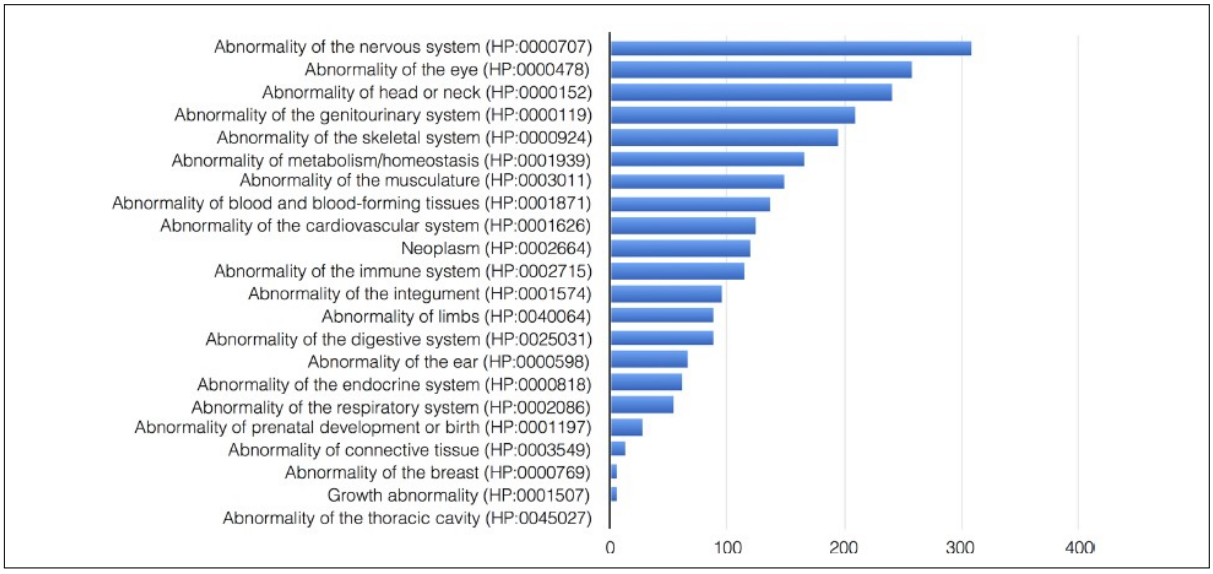

Figura 1.1: Distribución de las clases de HPO por categorías generales de anormalidades fenotípicas humanas. La imagen muestra el número de términos actuales añadidos por categoría. Imagen tomada de Köhler S, *et al.* 2017.

1.2. Objetivos y enfoque

Los objetivos generales del presente trabajo han sido los siguientes:

- Diseñar un pipeline de preprocesamiento para los informes de clínica que permita extraer su contenido en un formato que permita su manipulación.
- Extraer la terminología médica presente en el contenido de los informes mediante técnicas de text-mining.
- Realizar un mapeo automático de los conceptos médicos obtenidos a partir de los informes para su asociación a terminología HPO.
- Evaluar el impacto de la implantación algorítmica de Exomiser en el flujo analítico de NIMGenetics, comparando los resultados obtenidos tras utilizar los HPO procedentes del mapeo automático y los HPO asociados por un facultativo médico de la empresa.

1.3. Metodología y plan de trabajo

La programación realizada en el presente trabajo se llevó a cabo mediante el uso de la aplicación web *Jupyter Notebook* del paquete de distribución *Anaconda 3* y el entorno de desarrollo integrado *PyCharm* específico para el lenguaje Python.

El código desarrollado está subido en un repositorio *github* con el nombre *HPOMining* cuya dirección web se indica a continuación: https://github.com/jmiguel792/HPOMining.

El repositorio contiene un archivo *nimgenetics.py* que integra todas las funciones desarrolladas en el trabajo, las cuales son ejecutadas de forma secuencial en el código principal denominado *main_ code*. Este código corresponde con un *notebook* con extensión .ipynb desarrollado con Jupyter que utiliza las funciones del archivo .py para realizar los distintos procedimientos llevado a cabo a lo largo de este trabajo.

1.3.1. Plan de impulso de las tecnologías del lenguaje

En el presente trabajo se ha colaborado con el grupo de trabajo del Dr. Martin Krallinger del BSC que participa en el plan TL. El plan TL tiene como objetivo fomentar el desarrollo del procesamiento del lenguaje natural (NLP), la traducción automática y los sistemas conversacionales en la lengua española y lenguas cooficiales.

Las herramientas que actualmente están siendo desarrolladas por el plan TL se encuentran subidas en el siguiente enlace github: https://github.com/PlanTL-SANIDAD.

En este proyecto se ha utilizado la herramienta CUTEXT perteneciente al plan TL. Esta herramienta tiene como funcionalidad principal la extracción de terminología médica directamente de textos escritos. Además, permite extraer términos a partir de textos en distintos lenguajes como Inglés, Castellano, Gallego y Catalán. En este trabajo se utilizó la herramienta para la extracción de terminología médica a partir de informes cuyo contenido estaba relacionado con el diagnóstico clínico de pacientes con enfermedades genéticas raras.

1.3.2. Plataforma Health29

Health29 es una plataforma de índole innovadora desarrollada por la *Foundation29*. Esta organización sin ánimo de lucro está implicada en el enriquecimiento de *Health29* por medio del tratamiento de datos de pacientes. El objetivo final de *Health29* es facilitar y agilizar el diagnóstico del paciente para los médicos en los casos más complejos, donde la disponibilidad de datos respecto de la enfermedad sea muy reducida o se desconozcan los patrones de desarrollo.

La plataforma combina distintas herramientas públicas indicadas a continuación:

- *Exomiser*: herramienta utilizada para priorizar un conjunto de genes en relación con el diagnóstico clínico del fenotipo del paciente.
- *Phenomizer*: herramienta utilizada para identificar síntomas relacionados en función de los genes potencialmente asociados al fenotipo del paciente.
- OMIM y Orphanet: Son bases de datos públicas que albergan todo un catálogo de genes implicados en desórdenes genéticos. La información genética relacionada con enfermedades raras contenida en ambas bases de datos es fundamental para la correcta ejecución de la herramienta.

Por otro lado, la plataforma necesita incorporar información fenotípica del paciente aportada en forma de términos HPO. Por ello, cuanto mejor enriquecido esté el diagnóstico clínico del paciente, la probabilidad de obtener resultados significativos aumentará de forma considerable.

1.3.3. Casos de estudio seleccionados por NIMGenetics

Para realizar las correspondientes pruebas se han utilizado diez casos de estudio previamente seleccionados por el departamento médico de NIMGenetics. En función de las características del caso se aplicó el enfoque analítico WES más apropiado.

Las aproximaciones analíticas utilizadas según en qué caso fueron las siguientes: ExoNIM Dirigido, ExoNIM Clínico y ExoNIM Trío.

- ExoNIM Dirigido: análisis del exoma basado en la selección de genes asociados al fenotipo del paciente. Tras la secuenciación de 19.000 genes del paciente, se analizan aquellos asociados al cuadro clínico.
- ExoNIM Clínico: análisis del exoma focalizado en genes con fenotipo OMIM y/o con información clínica asociada previamente.
- ExoNIM Trío: análisis similar al clínico, pero que también implica la secuenciación de los 19.000 genes de los padres, lo que permite determinar rápidamente la procedencia de la variante o si se trata de una variante de *novo*.

Estos casos fueron imprescindibles para la comprobación y testeo de los procedimientos desarrollados a nivel de programación y los resultados obtenidos mediante la plataforma *Health29*.

2

The Human Phenotype Ontology. Estado del arte

2.1. Introducción

La terminología HPO surge como posible soporte para estandarizar el lenguaje humano no estructurado del médico que realiza el diagnóstico de la enfermedad. El hecho de asociar una terminología específica para describir las anormalidades fenotípicas del paciente, permite utilizar algoritmos como exomiser que son capaces de priorizar una lista de variantes genéticas entre las cuales pueda encontrarse la causante del fenotipo patológico. No obstante, actualmente el uso de HPO no está del todo integrado en el ámbito clínico, por lo que la identificación de la variante causal es dependiente del analista, lo que se traduce en términos muy costosos a nivel de tiempo y posibles errores humanos.

2.2. El Proyecto HPO

El proyecto HPO consiste en el desarrollo de una ontología médica compuesta por términos organizados jerárquicamente asociados a patologías fenotípicas humanas. Esta terminología se utiliza en diferentes ámbitos: investigación traslacional, diagnóstico diferencial y aplicación algorítmica en biología computacional [15]. En este último, los esfuerzos están dirigidos a potenciar la medicina personalizada utilizando técnicas de computación que incorporan los conceptos relacionados con las anormalidades fenotípicas observadas en el individuo, para así favorecer el enriquecimiento de HPO a nivel de diagnóstico [18].

A nivel técnico, HPO es actualmente una ontología de más de 13.000 términos cuya estructura está organizada como un grafo acíclico (DAG) donde las conexiones implican relaciones entre nodos o entidades que van de menor a mayor especificidad. Este tipo de representación jerárquica es útil porque cada entidad está localizada en una ruta ontológica específica, la cual se corresponde con el dominio conceptual de un término en particular [19] **(Figura 2.1)**.

La terminología HPO se adecua perfectamente a los objetivos de este proyecto debido a que durante su primera década de desarrollo (2007-2017), la base de su construcción han sido las enfermedades raras de tipo mendeliano. Las primeras descripciones de las versiones iniciales de HPO fueron elaboradas a partir de la base de datos OMIM. En años posteriores, la terminología HPO se ha ido enriqueciendo gracias al trabajo de clínicos profesionales que comenzaron a

colaborar extendiendo las descripciones terminológicas en áreas médicas más específicas como la cardiología o la inmunología [19].

Más recientemente, el proyecto HPO se ha enfocado en la expansión de otras áreas de la medicina relativas a enfermedades más comunes. Los resultados de esta nueva iniciativa involucran técnicas de text-mining muy elaboradas que son útiles para explotar y recuperar información de bases de datos como Pubmed del NCBI. Actualmente más de 132.000 anotaciones de términos HPO son resultado del trabajo de estas colaboraciones, a los cuales le restan los esfuerzos de curación por parte de profesionales médicos encargados de depurar las descripciones sintomatológicas de estos nuevos conceptos y su posterior aplicación en diagnóstico clínico [20].

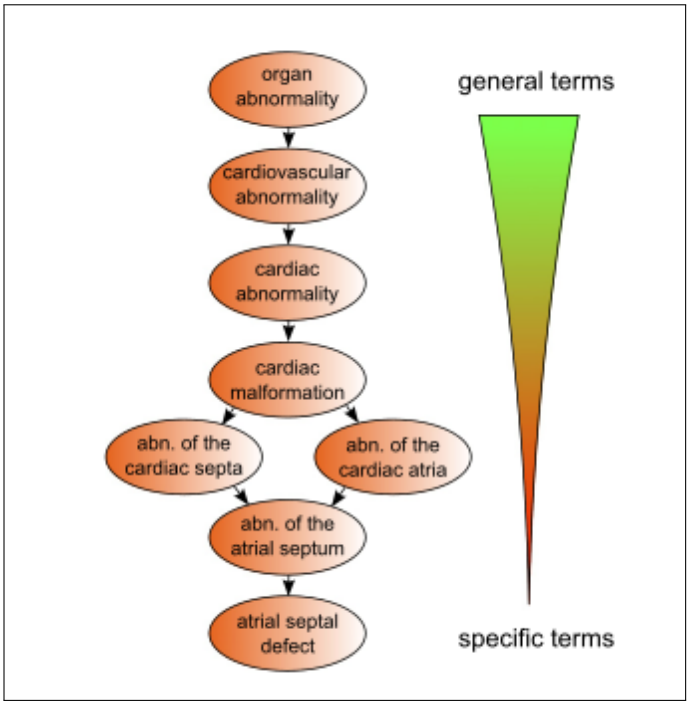

Figura 2.1: Ejemplo de representación de términos HPO como un DAG. Las entidades más generales (nodos parentales) involucran subclases cada vez más específicas dentro de las relaciones entre conceptos. En este caso en particular, el término HPO *abn. of the atrial septum* es un nodo hijo de *abn. of the cardiac septa* y *abn. of the cardiac atria*. A su vez, este término actúa como nodo padre de *atrial septal defect* que sería el HPO más específico de toda la ruta ontológica. Imagen tomada de Köhler S, *et al.* 2009.

2.3. Términos HPO

Cada término HPO describe una anormalidad de tipo clínico. Además, cada término pertenece de forma íntegra a una de las cinco subontologías independientes de la base de datos HPO [15]. *Phenotypic abnormality* es la ontología principal e integra todas las anormalidades fenotípicas descritas. *Mode of inheritance*, es una ontología relativamente pequeña que incluye términos relacionados con trastornos genéticos que cumplen patrones de herencia mendelianos. *Frecuency* es una ontología desarrollada con la colaboración de Orphanet en la cual se indica la frecuencia de aparición de características clínicas en los individuos. Por último, las subontologías *Clinical Course* y especialmente *Clinical Modifier*, son ontologías diseñadas para mejorar la precisión de las descripciones de anormalidades fenotípicas incluidas en la ontología principal [20]. *Clinical Course*, aporta información relacionada con patrones de tiempo y mortalidad, además del inicio y ritmo de progresión de la enfermedad. Por otra parte, *Clinical Modifier*, es una subontología que aporta términos HPO adicionales organizados en diversas categorías tales como: velocidad

de progresión, agravantes y factores manifestantes de la enfermedad, localización y severidad de la patología del individuo [21] [22] **(Figura 2.2)**.

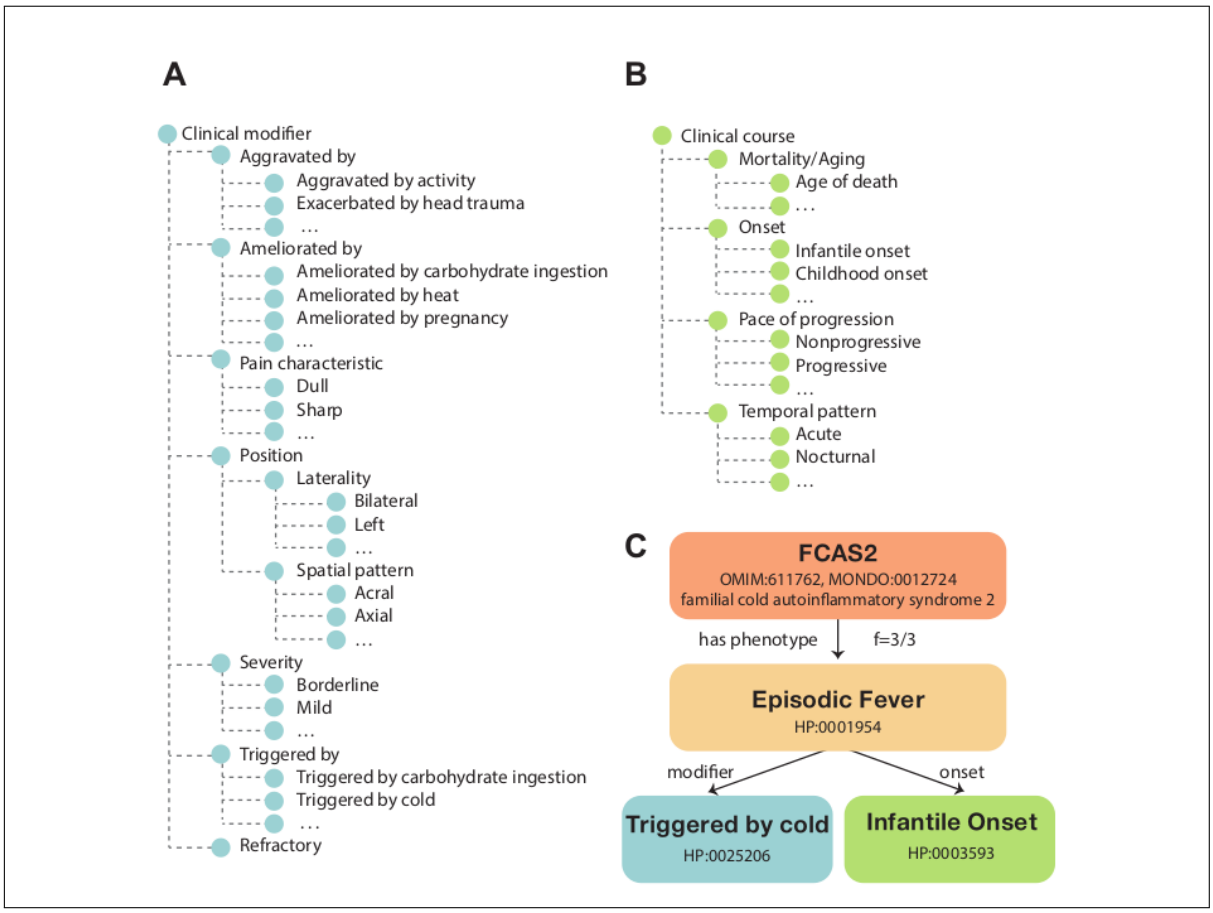

Figura 2.2: Visión general de las subontologías *Clinical Modifier* **(A)** y *Clinical Course* **(B)**. Ambas subontologías aportan términos HPO adicionales que sirven para enriquecer el significado de términos ya existentes. En la parte inferior derecha de la imagen **(C)** se visualiza un ejemplo de anotación de la enfermedad FCAS2. Los individuos que padecieron esta enfermedad fueron diagnosticados con episodios febriles (HP:0001954) en la infancia (HP:0003593) desencadenados por una exposición prolongada al frío (HP:0025206). Imagen tomada de Köhler S, *et al.* 2019.

Por otro lado, cada término se caracteriza por tener un identificador único (ID) y un nombre o etiqueta descriptivos. Además, la mayoría de los términos poseen descripciones adjuntas, las cuales especifican información relativa a la patología con la que el término HPO mantiene una relación directa. En la web de la base de datos HPO https://hpo.jax.org/app/ es posible realizar búsquedas de los términos que se desee obtener información en particular.

Un ejemplo de búsqueda podría ser el término *Limbal Dermoid* cuyo ID es HP:0001140. Al introducir el término en la base de datos, se obtiene información relativa al término buscado, la cual incluye una descripción y un conjunto de sinónimos. Además, es posible visualizar tanto las enfermedades como los genes asociados al mismo, sin olvidar que también es reproducible la jerarquía ontológica donde se localiza. Esta información procede en su mayoría de bases de datos relacionadas con enfermedades raras, como pueden ser OMIM y Orphanet. En este caso el término está integrado jerárquicamente dentro de una ruta ontológica relacionada con una anormalidad fenotípica de la córnea y esclerótica ocular y anomalías en la morfología del tejido conectivo de revestimiento **(Figura 2.3)**.

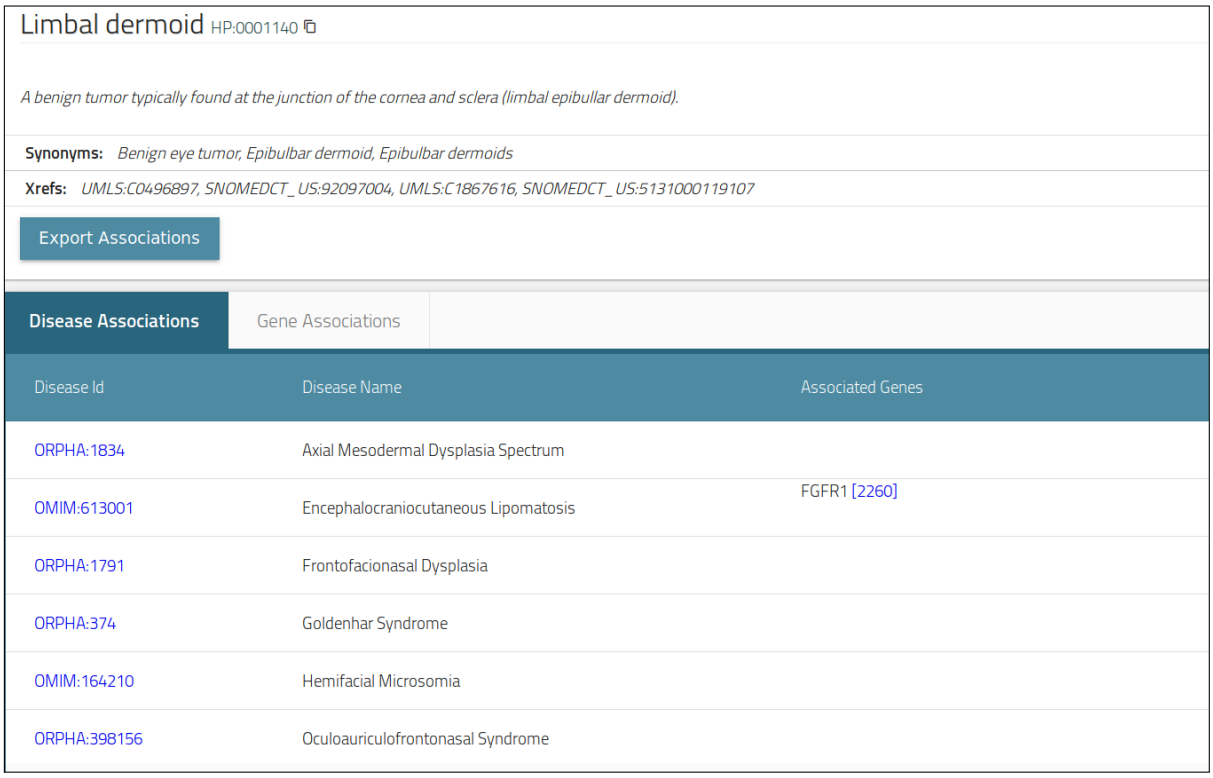

Figura 2.3: Captura de pantalla de la base de datos de HPO: https://hpo.jax.org/app/browse/term/HP:0001140 referente a la búsqueda del término *Limbal dermoid* con ID HP:0001140. Se pueden observar las características típicas que acompañan a cada término HPO: identificador, nombre, descripción, referencias, enfermedades y genes asociados.

2.4. Anotaciones

El dominio principal de las anotaciones de términos HPO han sido las enfermedades raras desde el inicio de la constitución del proyecto, donde han participado en su elaboración instituciones tan importantes como OMIM, Orphanet o DECIPHER. No obstante, como ya se ha comentado en el apartado 2.2, los recientes avances en medicina personalizada han impulsado que las anotaciones se desplieguen hacia otros campos de la medicina y, en este contexto, que se produzcan actualizaciones e incorporaciones constantes de nuevos términos a la base de datos de HPO [15] [21].

Estas nuevas anotaciones no sólo proceden de la elaboración de términos mediante información recuperada de Pubmed, sino que también están involucradas otras fuentes de información que han ido emergiendo debido a los nuevos objetivos del proyecto. Entre ellas destaca la ontología MONDO, la cual incorpora un nuevo conjunto de términos HPO que han sido elaborados a partir de unificar información procedente de OMIM y Orphanet, además de otras bases de datos como *The Human Metabolome Database* (HMDB) [21].

2.5. Integración de HPO

El proyecto HPO es el producto más importante perteneciente a la iniciativa Monarch. Esta plataforma se origina internacionalmente como un consorcio dedicado a la integración semántica de datos fenotípicos procedentes de fuentes de información relacionadas con la enfermedad humana y de distintas bases de datos de modelos de organismos animales. Su principal objetivo

es unificar todo este tipo de información para mejorar la investigación traslacional en el ámbito biomédico [23].

El proyecto adquiere vital importancia cuando existen enfermedades cuya base genética es desconocida o está asociada a diversas mutaciones en múltiples genes. En estas situaciones la calidad y consistencia de los datos integrados es un factor importante para la capacidad de computar fenotipos específicos de la enfermedad, los cuales pueden proporcionar información relevante respecto de la complejidad genética subyacente [24] **(Figura 2.4)**.

Conforme el proyecto ha crecido en términos de alcance, HPO se ha convertido en uno de los componentes principales de la iniciativa Monarch, debido a que actúa como un puente entre los datos genéticos disponibles y la elaboración de perfiles fenotípicos de enfermedades humanas [15]. En este contexto, HPO es un completo éxito en cuanto a la integración de información (el principal objetivo de la iniciativa) entre distintos campos científicos y fuentes de información, pero lamentablemente poco establecido dentro de la comunidad científica para el desarrollo de investigaciones de tipo traslacional o para tomar decisiones a nivel de diagnóstico clínico [23].

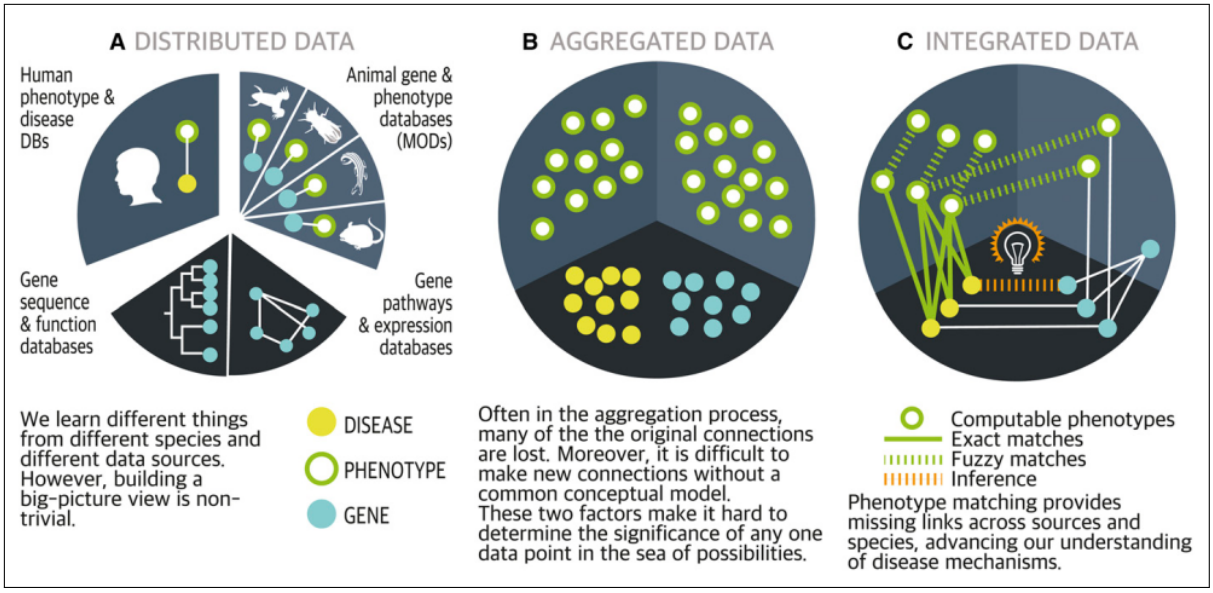

Figura 2.4: Protocolo de integración de la información fenotípica disponible. **A)** Heterogeneidad de datos sin unificar con distinta procedencia (humana, animal, etc) y diversa escala biológica (genética, fenotípica y enfermedad). **B)** Proceso de unificación y agregación de datos fenotípicos adicionales que proporcionan información frente a casos con base genética desconocida. Este proceso apuesta por perder información cuantitativa para **C)** computar fenotipos que faciliten la interpretación de mecanismos subyacentes a la enfermedad. Este último paso involucra la integración global de la información inicial, cuyo origen era muy diverso, para así proporcionar nuevas perspectivas respecto de las relaciones generadas entre diferentes fuentes de información y entre distintas especies. Imagen tomada de McMurry JA, *et al.* 2016.

2.6. Interoperatividad HPO

Los fenotipos clínicos pueden ser descritos utilizando múltiples terminologías, entre ellas destacan HPO y otras como *The Standardized Nomenclature of Medicine Clinical Terms* o *Unified Medical Language System*, más conocidas como SNOMED CT y UMLS respectivamente, cuya terminología médica está estandarizada a nivel global. Lo que diferencia a HPO de las otras terminologías radica en que actualmente su uso está muy fomentado en los grandes proyectos de los institutos de salud internacionales (NIH), como por ejemplo ClinVar, además de por las

capacidades funcionales que ofrece para potenciar el diagnóstico genético mediante su aplicación en algoritmos de priorización de variantes como exomiser.

La interoperatividad de los datos fenotípicos anotados entre terminologías juega un papel muy importante en el ámbito de la investigación traslacional [25]. Por ejemplo, algunos casos de selección de cohortes establecen sus patrones de reclutamiento de pacientes en función del fenotipo que presentan, el cual es determinado específicamente utilizando como referencia la combinación de terminologías como HPO y SNOMED CT [26].

Se ha registrado que HPO mapea UMLS en un 54 % y tan solo un 30 % en SNOMED [27]. UMLS, por su parte, como sistema de integración de terminologías estándar [28], puede facilitar el mapeo entre HPO y SNOMED, siendo posible de forma parcial o completa en función de si ambas terminologías comparten conceptos entre sí [26] **(Figura 2.5)**.

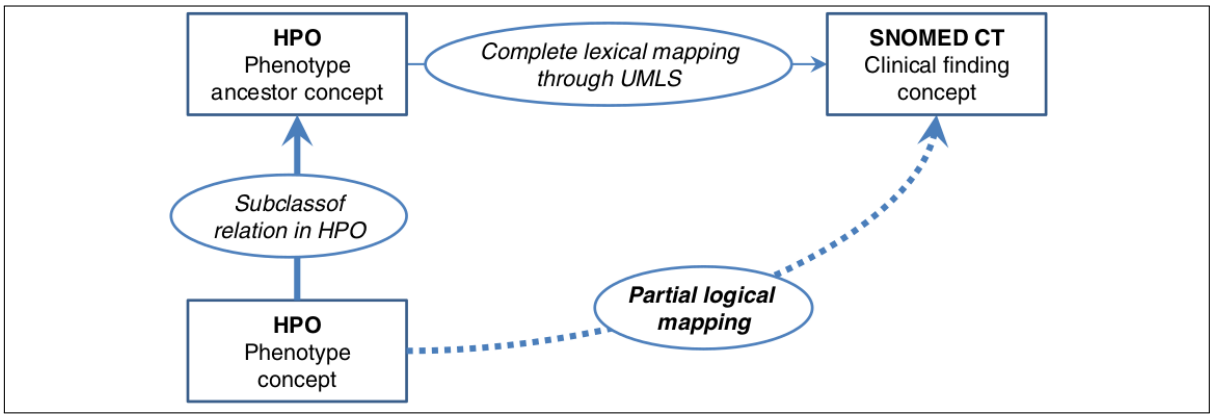

Figura 2.5: Representación de un mapeo de tipo parcial y completo entre las terminologías HPO y SNOMED a través de UMLS. Imagen tomada de Dhombres F, *et al.* 2016.

Por otro lado, HPO debe su estructura a dos tipo de formato comúnmente utilizados para el desarrollo de ontologías terminológicas: *Web Ontology Language* (OWL) y *Open Biomedical Ontology* (OBO). La versión OWL de HPO contiene algunas características que la versión OBO no incluye, como son las definiciones lógicas de los términos HPO [21]. Por ejemplo, la definición lógica del término HPO *Astigmatism* seguiría la siguiente estructura:

'has part' some

('asymmetrically curved'

and ('inheres in' some cornea)

and ('has modifier' some abnormal))

La representación de los términos mediante estas definiciones permite facilitar los procesos de computación de datos fenotípicos para interoperar entre otras ontologías cuya terminología incluye información fenotípica de otras especies (apartado 2.5). Estas ontologías son muy diversas: Uberon, Gene Ontology, The Cell Ontology o Phenotype, Attribute and Trait Ontology (PATO), etc. Además, estas definiciones lógicas actúan como punto de control para inferir nuevas clasificaciones ontológicas si el análisis fenotípico establece nuevas relaciones entre distintas especies [29].

La versión OBO de HPO es más sencilla de interpretar. El formato OBO es ampliamente utilizado para desarrollar diversas bio-ontologías [30] **(Figura 2.6)**. A diferencia de OWL, la versión OBO de HPO organiza el contenido de los términos en bloques. Cada bloque incluye los elementos característicos de los términos HPO (apartado 2.3), además de la posibilidad de albergar un ID alternativo, un comentario (aparte de la descripción correspondiente) y el tipo de relación que existe con otros HPO, que se indica mediante el formato *is_a*. Los términos

HPO indicados como *is_a* son entidades padre de la categoría ontológica a la que pertenece el HPO principal del bloque, por lo que dicho HPO es una entidad hija que aporta información más específica respecto de la patología asociada a este término en particular. Por ejemplo, la representación OBO del término *Functional abn. of the bladder* sería de la siguiente forma:

[Term]

id: HP:0000009

name: Functional abnormality of the bladder

alt_id: HP:0004424

alt_id: HP:0008731

def: "Dysfunction of the urinary bladder."

synonym: "Poor bladder function"

xref: UMLS:C3806583

is_a: HP:0000014 ! Abnormality of the bladder

Este tipo de representación, aparte de ser más interpretable, facilita también la interoperatividad entre clases de términos que comparten un nicho de información relacionado. Por ejemplo, el proceso de replicación del DNA como área de conocimiento, debería resultar en la inferencia de términos que fueran comunes entre distintas especies y que estuvieran relacionados con este tipo de proceso [30].

Domain	Name (Abbreviation, Reference)	Downloaded file (relative to http://purl.obolibrary.org/obo/)
biochemistry	Chemical Entities of Biological Interest (ChEBI[29])	chebi.obo
	Gene Ontology (GO[30])	go.obo
proteins	Protein Ontology (PRO[31])	pr.obo
cell types	Cell Ontology (CL[32])	cl.obo
anatomy	Foundational Model of Anatomy (FMA[33])	fma.obo
	Spatial Ontology (BSPO[-])	bspo.obo
	Mouse adult gross anatomy (MA[34])	ma.obo
	Zebrafish anatomy and development (ZFA[35])	zfa.ob
	Multi-species anatomy (UBERON[36])	uberon.obo
phenotype	Phenotype, Attribute and Trait Ontology (PATO[22])	pato.obo
	Mouse Pathology (MPATH[37])	mpath.obo
	Mammalian Phenotype Ontology (MPO[17])	mp.obo
	Human Phenotype Ontology (HPO[18])	hp.obo
	Neuro Behavior Ontology (NBO[38])	nbo.obo

Figura 2.6: Ontologías desarrolladas mediante el formato OBO entre las que se encuentra HPO. Este tipo de representación permite organizar los términos en distintas categorías o clases para facilitar la clasificación de las anormalidades fenotípicas descritas. Estas bio-ontologías incluyen terminología referente a diferentes especies como zebrafish, ratón o humano. Imagen tomada de Köhler S, *et al.* 2013.

3

Sistema, Diseño y Desarrollo

3.1. Introducción

Como se ha comentado en la introducción del presente trabajo, NIMGenetics tiene a su disposición un gran número de informes clínicos de pacientes cuyo contenido se presenta en lenguaje natural. Por ello, se han utilizado técnicas de text-mining con el objetivo de extraer terminología médica procedente del texto de los documentos. Concretamente este proceso involucra técnicas NLP que son necesarias para la extracción, análisis y correlación de términos cuya relevancia sea significativa en función del diagnóstico establecido por el profesional médico correspondiente.

La extracción de los conceptos médicos de los informes es necesaria para su asociación de los correspondientes términos HPO. Sin embargo, para alcanzar este punto, un conjunto de procedimientos relativos al mapeo, preparación y preprocesamiento de los informes son necesarios para la correcta consecución de los procesos comentados en este apartado.

3.2. Extracción de informes clínicos de Gestlab

Los informes de clínica de pacientes que dispone NIMGenetics se encuentran almacenados en su base de datos: *Gestlab*. La carpeta que contiene todos los informes ocupa un espacio de aproximadamente 300 GB (*Gestlab DOC*), la cual se copió directamente en el local del ordenador de trabajo. Esta carpeta está organizada como un árbol de carpetas donde hay documentos en diversos formatos. El formato de los informes de clínica es PDF.

3.2.1. Mapeo de los informes de clínica

Se desarrolló el programa *load_files* para el mapeo de todos los PDF incluidos en el árbol de carpetas, obviando los documentos almacenados en otro formato distinto. El programa permite realizar dos tipos de mapeo: a) un mapeo donde se obtiene una lista con el path absoluto para todos los informes y b) un mapeo individual para obtener una lista con los informes individualizados con su identificador correspondiente. Ambas listas se utilizarán más adelante a lo largo del desarrollo global del código principal **(MS-A)**.

3.2.2. Obtención del identificador real de los informes

Hay que destacar que el identificador de los informes se encuentra codificado en hexadecimal cumpliendo la normativa LOPD, lo cual impide conocer el identificador real del paciente. El esquema de nomenclatura por el cual se rigen los informes es complejo. Un ejemplo del identificador codificado en hexadecimal de un informe sería de la siguiente forma: 0100_0015C98_001B963. Como se puede observar, se pueden visualizar tres componentes separados por guiones bajos. El segundo componente, en este caso 0015C98, es de los tres el que proporciona información relevante respecto del identificador real del informe.

La transformación del identificador de hexadecimal a decimal, se realizó iterando sobre la lista obtenida mediante mapeo individual. De esta forma se generó una nueva lista que almacenó todos los identificadores en el formato apropiado para realizar los sucesivos procedimientos a partir de este punto **(Figura 3.1)**.

Este paso es necesario para volcar esta nueva lista sobre la función *getIDs*, desarrollada específicamente para realizar una consulta SQL sobre Gestlab con el objetivo de obtener los identificadores reales. Esta consulta se ejecuta sobre la tabla *petición* de Gestlab, donde el identificador decimal obtenido en el paso anterior es utilizado para establecer la búsqueda del identificador real. El resultado final del proceso es la obtención de una nueva lista que alberga el identificador real de los informes pertenecientes a los pacientes que NIMGenetics dispone en sus archivos **(MS-A)**.

```
id_peticion = []
for pdf in l_pdf:
    # split por guion bajo
    s_pdf = pdf.split('_')
    # transformación del segundo componente del esquema hexadecimal del identificador
    hex_id = int(s_pdf[1],16)
    id_peticion.append(hex_id)
```

Figura 3.1: Fragmento de código perteneciente al *main_code* donde se observa la lista generada después de transformar el segundo componente de hexadecimal a decimal. La nueva lista integra los identificadores que serán utilizados para realizar consultas SQL a Gestlab para obtener los identificadores reales de los informes de clínica.

3.2.3. Almacenamiento de los informes con su identificador correcto

La lista obtenida se caracteriza por ser una lista de listas de identificadores reales y elementos vacíos. Estos últimos se producen cuando la consulta SQL no encuentra el identificador correspondiente en Gestlab. Estos casos son el resultado de la eliminación o cambio del nombre del identificador del informe, por lo que se obtiene un elemento que no aporta información. El hecho de que algunos elementos sean vacíos dificulta el siguiente paso, el cual implica guardar los informes con su identificador correcto. Para llevar a cabo este procedimiento, se genera una nueva lista (*l_names*) que almacena todos los identificadores y componentes vacíos con el nombre identificativo *sin_id*. La ejecución de este proceso utiliza un sistema de manejo de excepciones que permite añadir a *l_names* ambos elementos **(Figura 3.2)**.

El paso anterior es importante porque *l_names* será utilizada en la ejecución del siguiente programa, el cual está diseñado expresamente para el almacenamiento de los informes con su identificador correcto en la nueva carpeta denominada *Resultados*. Por otro lado, la estructura de los informes se caracteriza por presentar tres componentes: a) el año de publicación del informe (dos cifras); b) una letra N seguida de otra letra indicativa del tipo de informe y c) un número referente al orden de publicación del mismo. Por tanto, siguiendo esta estructura, un ejemplo del

identificador de un informe se representaría de la siguiente forma: 17NR1527. Este identificador significa que el informe 1527 del año 2017 pertenece al tipo de informe NR. Respecto de los tipos de informe, se extrajeron los informes distinguidos por ser del tipo NR, NG y NS, donde las letras que preceden a N simbolizan el departamento de Bioinformática y NGS.

Teniendo en cuenta las premisas anteriores, las funciones *identifiedPDF* y *emptyPDF* fueron desarrolladas para extraer y almacenar los PDF en distintos directorios dentro de la carpeta *Resultados*. El primero se caracteriza por filtrar el tipo de informe utilizando la función *search* del módulo *re* como método de búsqueda. A continuación, el módulo *shutil* permite copiar el informe con el nombre correcto en el nuevo directorio. Este programa genera tres carpetas en *Resultados*: a) *Identificados*: informes que cumplen el patrón requerido (NR, NG, NS). b) *Repetidos*: informes cuyo nombre es idéntico. En general estos informes son ampliaciones del original que aportan más información clínica. c) *Descartados*: informes que no cumplen el patrón establecido (NT, 8G). Por otro lado, el segundo programa genera una carpeta nombrada como *sin_identificar*, la cual almacena todos los informes cuyo identificador no ha sido recuperado de la base de datos y, por tanto, se guardarán con el identificador hexadecimal original **(MS-A)**.

```
l_names = []
for row in r_id:
    try:
        # Primer bloque: extraer identificadores que cumplan la siguiente estructura
        l_names.append(row[0][0])
    except:
        # Segundo bloque: identificadores que devuelven elementos vacios
        # Estos informes se almacenarán con su identificador hexadecimal original
        s_row = str(row)
        if s_row == '[]':
            # Almacenamos en l_names los identificadores reemplazados con el literal 'sin_id'
            sr_row = s_row.replace('[]', 'sin_id')
            l_names.append(sr_row)
```

Figura 3.2: Fragmento de código perteneciente al main_code que representa la generación de la lista *l_names* al iterar sobre la lista *r_id* procedente de la ejecución de *getIDs*. La nueva lista *l_names* alberga tanto identificadores reales como elementos vacíos.

3.3. Pasos previos a la extracción de contenido

NIMGenetics escanea los informes recibidos directamente en formato PDF. El resultado es una imagen escaneada a la cual no es posible aplicar librerías de Python que sean capaces de extraer el texto a partir de este formato, por lo que otro tipo de estrategia se llevó a cabo con el objetivo último de extraer terminología. En este caso, la función *pdfToImage* fue diseñada para transformar la imagen en PDF a una imagen en formato PNG, la cual es susceptible de aplicar técnicas OCR de Python con el fin de generar texto plano.

3.3.1. Generación de imágenes a partir de los informes de clínica

Para llevar a cabo este paso, utilizaremos los informes extraídos en la carpeta *Identificados*, la cual incluye aquellos informes que cumplieron el patrón de selección. Para ejecutar la función *pdfToImage*, antes es necesario preparar el path absoluto de los informes por medio de la función absolutePath, la cual es utilizada en el trabajo en procedimientos similares **(Figura 3.3)**.

La correcta ejecución del programa genera una carpeta de imágenes dentro de *Resultados* junto con el resto de directorios. Esta carpeta alberga todas las imágenes correspondientes a cada informe, generándose tantas imágenes como páginas tenga el PDF a transformar. Esta función

utiliza la librería de Python *Wand* para la conversión de los informes a formato PNG, siendo estrictamente necesario eliminar las imágenes mediante la función *destroy* de la librería después de haber sido procesadas, para impedir su almacenamiento en memoria y así evitar la detención del programa por saturación de la misma. Para solucionar este tipo de errores, el código del programa alberga un sistema de manejo de excepciones que impide su detenimiento. Este hecho también podía producirse debido a que algunos informes eran detectados como erróneos por ocupar un espacio de 0 Mb, lo cual indicaba que su formato era inválido. Además, el sistema también permitía proseguir con la ejecución en los casos en que el informe estuviera encriptados con una contraseña.

Por último, como ya se ha comentado, las imágenes fueron guardadas en formato PNG, utilizando el mismo nombre del informe y añadiendo a continuación la palabra *image* junto con el número de página correspondiente al informe **(MS-A)**.

```
def absolutePath(dpath):
    path = dpath
    l_path = os.listdir(path)
    l_abs = []
    for ele in l_path:
        l_abs.append(os.path.join(path, ele))

    return l_abs

# pdfPath contiene el path absoluto de todos los pdf
# este paso es necesario para ejecutar el siguiente programa: pdfToImage
pdfPath = nimgenetics.absolutePath(dpath=pathI)
```

Figura 3.3: Fragmento de código perteneciente al *main_ code* combinado con la función *absolutePath* en este caso adaptada al directorio *Identificados*. La lista *pdfPath* alberga el path absoluto de todos los informes que cumplen el patrón de selección.

3.3.2. Conversión del formato imagen a texto plano

Tras la obtención de imágenes, el siguiente paso implica la ejecución de la función *makeCorpus* la cual genera la carpeta *textos* que contiene todos los informes procesados en texto plano. El código del programa integra un bucle que toma bloques de imágenes por informe, generando dos listas en cada iteración: a) una lista del bloque de imágenes a procesar y b) una lista con el texto extraído de cada bloque.

Este proceso es posible porque antes de generar la lista de imágenes, el código selecciona el bloque de imágenes a partir de una lista de identificadores únicos. A continuación, la lista de textos se genera iterando sobre el bloque de imágenes seleccionado. Este paso incluye la ejecución de un pequeño programa denominado *ocrToImage*, el cual utiliza el módulo *image_ to_ string* de la librería *pytesseract* de Python que es capaz de transformar el contenido de cada imagen en texto plano. Por último, la lista final de textos se transforma en un string y se guarda en la carpeta generada en *Resultados*. Cuando el proceso finaliza, otra iteración comienza con el siguiente bloque de imágenes y finaliza por completo cuando todos los bloques han sido procesados **(MS-A)**.

El resultado final es la elaboración de un corpus que contiene el texto de todos los informes procesados. El siguiente paso involucra la extracción de conceptos médicos encontrados en los textos generados a partir de los informes de clínica, los cuales se asociarán a su terminología HPO correspondiente con el objetivo de elaborar un perfil fenotípico del paciente lo más detallado posible que potencie la probabilidad de éxito del diagnóstico en experimentos sucesivos.

3.4. Técnicas para la extracción de terminología médica

Los métodos para la extracción de terminología médica del texto de los informes comprenden dos estrategias distintas: a) la extracción mediante cutext (apartado 1.3.1) y b) la extracción mediante la técnica de búsqueda de conceptos *lookup*. La totalidad de los términos extraídos será utilizada para realizar un mapeo automático a términos HPO, mientras que una selección más pequeña de 500 términos será utilizada para realizar una asociación manual.

3.4.1. Implementación funcional de Cutext

Como ya se ha comentado previamente, cutext es una herramienta para extraer terminología médica cuya funcionalidad se aplicó a los textos generados a partir de los informes. Su aplicación garantiza en un alto porcentaje la extracción de conceptos de los textos que tengan relación con lenguaje médico y, por otra parte, la inclusión de algunos falsos positivos que deberán ser eliminados posteriormente.

Antes de ejecutar cutext, se requiere la instalación de *TreeTagger*. Este paquete permite generar anotaciones de textos por medio de procesos de lematización y Part-of-Speech (PoS). Cutext utiliza el anotador para la extracción de conceptos médicos de los textos. Por otro lado, cutext está implementado en java, por lo que el programa desarrollado para ejecutarlo fue adaptado a esta premisa. Además, como cutext está en proceso de testeo actualmente, inicialmente elimina el documento de texto que se utiliza como base para la extracción, lo cual implica que fuera necesario utilizar copias de los textos procesados.

La herramienta se organiza internamente en varios directorios. En la carpeta *in* se introduce el texto a procesar y después se genera un output denominado *terms_raw.txt* en la carpeta *out*, habiendo sido eliminado el documento de la carpeta anterior **(Figura 3.4)**. El programa *getCutextResults* desarrollado para esta parte del trabajo, genera copias de los textos y almacena el output en una carpeta dentro de *Resultados* denominada *cutext_results* con el nombre de los textos originales. En relación con la implementación java de cutext, el programa utiliza el archivo jar que viene con cutext para las sucesivas ejecuciones de cada documento de texto **(MS-A)**.

3.4.2. Estrategia Lookup

La estrategia lookup comparte con cutext el objetivo de la extracción de terminología de los textos. En este caso, el proceso de extracción depende del contenido de un diccionario de términos elaborado por el grupo de colaboración del BSC. Este diccionario consiste en un documento que alberga multitud de conceptos y términos médicos, de forma que si alguno de estos conceptos aparece en los textos, dicho término será extraído y almacenado en un nuevo corpus generado a partir de esta estrategia.

A diferencia de cutext, esta estrategia es mucho más costosa a nivel computacional, por lo que en este caso se realizó una subselección de los informes que contenían más información a nivel de clínica del paciente. Este hecho fue necesario porque sino la ejecución de la estrategia para los más de 20.000 informes requería de un tiempo de computación aproximado de varias semanas. Por ello, para este proceso únicamente se tuvieron en cuenta los informes identificados con el nombre *PRUEBA SOLICITADA: EXONIM*. Este tipo de informe se caracteriza por presentar información restringida al diagnóstico del paciente, sin albergar otro tipo información relativa a otras pruebas como arrays, FISH o Sanger, que son menos informativas a nivel de fenotipo.

Tras haber seleccionado los informes que cumplían las condiciones anteriores, los procedimientos que siguen a continuación involucran la carga del diccionario y la preparación de términos

que aparecen en los textos a nivel de frase. El hecho de que sea específicamente a nivel de frase es importante debido a que de esta forma los términos se puedan extraer íntegramente según como aparecen en la frase analizada, lo cual no sería posible si el lookup se ejecutase a nivel de palabra. Por ejemplo, no sería lo mismo la extracción del término diabetes (a nivel de palabra), que diabetes tipo II (a nivel de frase) el cual aporta más información tanto a nivel conceptual como a la hora de mapear los HPO correspondientes para dicho concepto.

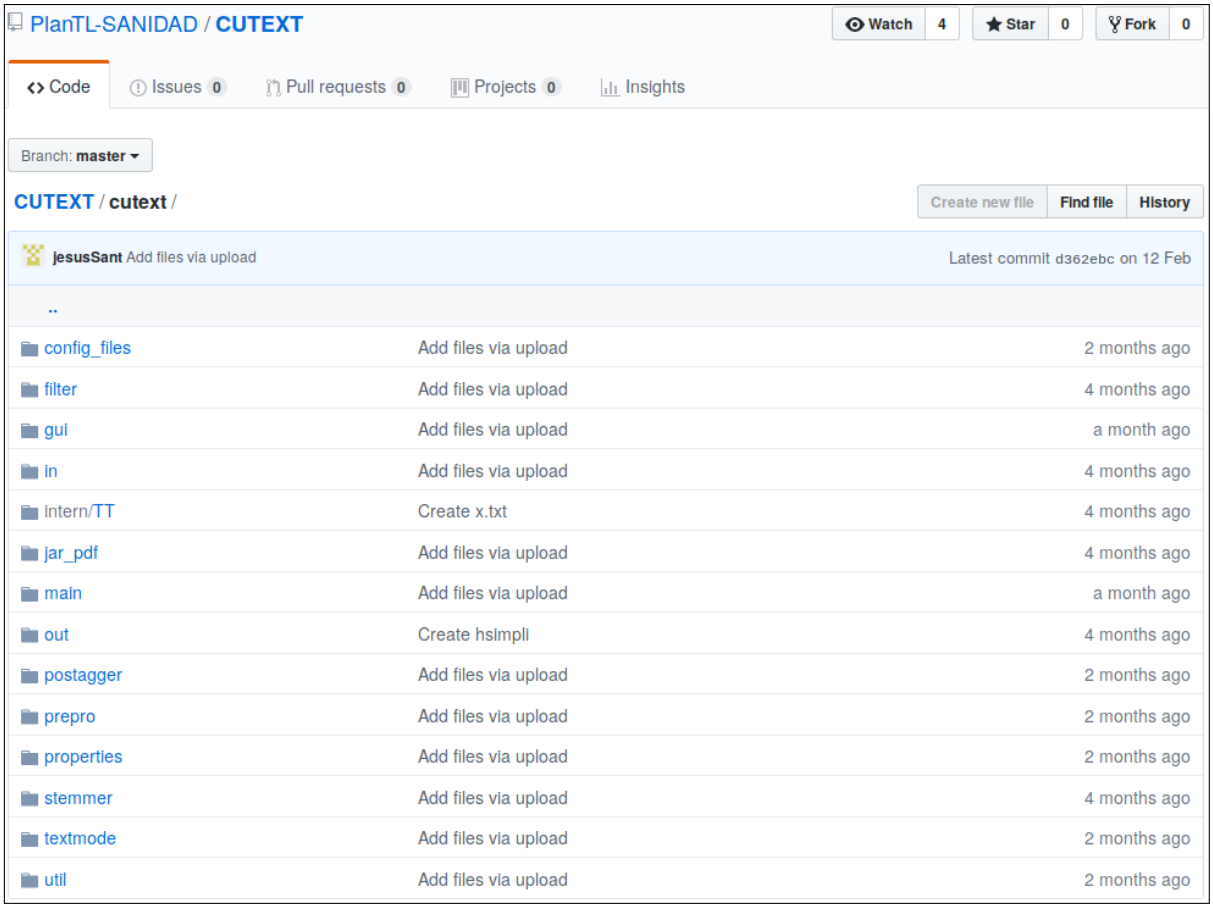

Figura 3.4: Captura de pantalla de https://github.com/PlanTL-SANIDAD/CUTEXT/tree/master/cutext donde se puede observar el repositorio github de CUTEXT del planTL que muestra la estructura interna de los directorios integrados en cutext incluidas las carpetas *in* y *out* citadas en este apartado.

4

Experimentos Realizados y Resultados

4.1. Introducción

En este capítulo se detalla la parte experimental realizada en el presente trabajo, la cual contempla diversos apartados: la aplicación y adaptación funcional de las técnicas de extracción sobre los textos de partida, el mapeo automático de los conceptos extraídos a HPO, la selección de un subconjunto para su asociación manual y la ejecución de exomiser y *Health29* utilizando la terminología HPO obtenida para describir el fenotipo patológico del paciente.

4.2. Análisis de los métodos de preprocesamiento

En este apartado se muestra el rendimiento de las funciones diseñadas para el preprocesamiento de los informes de clínica, donde se especifica el tiempo de ejecución de cada programa y el análisis inicial del corpus obtenido. Para el cálculo de los tiempo de ejecución, se utilizó el módulo *time* de Python **(Tabla 4.1)** y para examinar el contenido del corpus se utilizaron módulos integrados de la librería NLTK **(Tabla 4.2)**.

En la tabla 4.1 se puede observar que las funciones que requieren un mayor tiempo de ejecución son aquellas implicadas en la generación de imágenes y en la elaboración del corpus. En concreto, la construcción de imágenes implicó un gasto computacional más elevado debido a que la longitud de los informes no era uniforme. Este hecho explicaba que los tiempos de ejecución de la generación de textos fueran menores en relación con las imágenes, ya que el proceso era más rápido cuando los informes contenían pocas páginas, pero se hacía significativamente más lento cuando el número de páginas aumentaba.

Por su parte, la tabla 4.2 muestra un resumen analítico del corpus. Estos datos se calcularon utilizando un *notebook* en el cual se integraron fórmulas específicas para el cálculo de estas frecuencias. Este documento se puede visualizar en el repositorio github del presente trabajo: https://github.com/jmiguel792/HPOMining/blob/master/frecuencias_corpus.ipynb.

Se puede observar que 26.707 informes fueron procesados y que más de 23 millones de tokens conforman la totalidad del corpus. Sin embargo, la diversidad léxica que existe entre el número total de tokens y el número de tokens diferentes se restringe a un 3,2 %, por lo que se deduce que muchas palabras son similares en el contexto clínico de los informes.

Función	Tiempo de ejecución
load_files	5 min
getIDs	25 min
identifiedPDF	4,2 h
emptyPDF	20 min
pdfToImage	72 h
makeCorpus	60 h

Tabla 4.1: Tiempos de ejecución calculados para cada una de las funciones desarrolladas con el objetivo de realizar los procedimientos preestablecidos para el preprocesamiento de los informes de clínica.

Frecuencias Corpus	
Número total de documentos procesados	26.707
Número total de tokens	23.394.717
Número total de tokens diferentes	747.270
Número total de frases	1.860.776
Número total de párrafos	1.358.677
Media de tokens por informe	≈375
Media de frases por informe	≈70

Tabla 4.2: Datos analíticos relativos al corpus generado mediante la función *makeCorpus*.

4.3. Corpus elaborado mediante Cutext

La ejecución de cutext genera tantos documentos como textos se hayan procesado, por lo que la cantidad de archivos que cutext produce es similar a lista de textos originales. El output de cutext se caracteriza por presentar bloques de términos extraídos del texto original procesado, los cuales están precedidos por cuatro elementos informativos: lema del término, Cvalue, frecuencia y longitud del término **(Tabla 4.3)**.

Term: nevus melanocítico
Lema: nevus melanocit
CValue: 2.0
Frecuency: 1
Length: 19

Tabla 4.3: Representación del término *nevus melanocítico* extraído del informe clínico de un paciente. Se observan los cuatro elementos característicos en bloque que proporciona la ejecución de cutext.

A pesar de que este tipo de información podría resultar valiosa para la selección específica de términos en función del contexto del informe, se decidió extraer todos los términos de todos los bloques generados en cada output para no perder información. Este hecho se puede explicar poniendo de ejemplo la extracción de conceptos únicamente cuando su frecuencia en el texto era de 1. Cuando la extracción se realizaba de esta forma, efectivamente se obtenían los términos cuya frecuencia en el texto era de 1, pero, por otra parte, con el fin de buscar enriquecer el corpus con las palabras menos frecuentes (pero a su vez más informativas), se observó que también se perdían conceptos cuya frecuencia era mayor y que también resultaban informativos respecto del contexto clínico del informe. Por ello, se decidieron extraer todos los términos de cada bloque, aceptando que podían introducirse falsos positivos.

El programa *cutextCorpus* se diseñó para llevar a cabo el proceso explicado en el párrafo anterior. El programa utiliza como parámetro la lista completa de los resultados generados por cutext (apartado 3.4.1) e itera sobre ella para generar un corpus de términos que se almacena en *Resultados* dentro de un nuevo directorio denominado *Extracción* con el nombre de *cutext_ corpus*. Este corpus se caracteriza por ser un documento de texto que integra todos los términos de todos los textos procesados por cutext **(MS-A)**.

4.4. Corpus elaborado mediante la estrategia Lookup

Como se ha detallado anteriormente, la estrategia Lookup involucra la carga del diccionario elaborado por el grupo de colaboración del BSC y la preparación de la terminología de los informes, lo cual implica la integración y almacenamiento de los conceptos de ambas fuentes de información en listas diferentes. En esta situación, la función *lookup* está específicamente diseñada para recorrer las listas anteriores y extraer la terminología en común, obteniéndose un nuevo corpus que integra el vocabulario compartido entre el diccionario y los conceptos de los informes seleccionados. Este corpus es almacenado en el directorio *Extracción* junto con el corpus generado mediante cutext con el nombre de *corpus_ lookup_ procesado* **(MS-A)**.

4.5. Resumen de los métodos implementados: Cutext y Lookup

En este apartado se recopila el número de conceptos extraídos mediante cutext y la estrategia lookup, además de los tiempos de ejecución necesarios en ambos casos **(Tabla 4.4)**. Destacar que para el set de términos final fue necesario un proceso de curación de falsos positivos que inevitablemente se generaron durante la elaboración del corpus. Para ello se aplicaron una serie de filtros que eliminaban aquellos conceptos que podían desvirtuar el contenido del corpus.

Este filtrado omitía los términos que se repetían o que no alcanzaban una longitud mínima, además de suprimir aquellos que comenzaban con un elemento inválido, lo cual provocaba que ese término no aportase significado alguno como concepto médico. Sin embargo, a pesar de aplicar estos filtros, otro procedimiento adicional fue necesario, el cual implicaba realizar una curación estrictamente manual del corpus, para así asegurar una integridad y enriquecimiento adecuada para los sucesivos experimentos a realizar.

Técnica	**Función**	**T Ejec**	**Términos sin curar**	**Terminología única**
Cutext	*getCutextResults* *cutextCorpus*	20 h 30 min	18.235.562	1131
Lookup	*lookup*	5 días	27.398.941	4186

Tabla 4.4: Terminología global obtenida tras ejecutar cutext y la estrategia lookup.

Se puede observar una diferencia clara entre el número de términos extraídos sin establecer ningún proceso de curación y tras aplicar curación manual. La estrategia lookup requiere de un tiempo de computación muy superior a cutext obteniéndose un mayor número de términos a partir de los informes. No obstante, este hecho no implica que los conceptos obtenidos mediante cutext sean menos informativos o que la terminología extraída fuera similar, puesto que tan sólo se compartían 310 términos respecto de los 5317 totales. Este conjunto total de términos será utilizado para realizar el mapeo automático a términos HPO.

4.6. Elaboración de un corpus global

Se desarrolló la función *mergeCorpus* para elaborar un corpus global a partir de los corpus construidos mediante cutext y la estrategia Lookup. Este corpus se caracteriza por contener el vocabulario común y no común extraído por ambas técnicas. Este hecho implica que si existen dos conceptos similares entre los corpus de cutext y lookup, sólo se integrará ese término una única vez en el corpus global, además de los conceptos no compartidos. Este nuevo set de vocabulario se almacena en el directorio *Extracción* con el nombre de *corpus_ global* **(MS-A)**.

Este corpus fue utilizado para realizar el mapeo automático a HPO y para seleccionar un subconjunto de términos del mismo para llevar a cabo una asociación de tipo manual. Por otro lado, permitió construir una matriz de distribución terminológica que hiciera posible conocer el número de veces que aparecía cada término en los informes de clínica para así tener controlado el fenotipo de cada paciente a nivel conceptual.

4.6.1. Matriz de distribución de terminología médica

La elaboración de un corpus global nos permitió, aparte de conocer los términos extraídos comunes entre ambas estrategias, analizar la distribución de los términos a nivel global. En este apartado, los procesos llevados a cabo se centran en la construcción de una matriz que represente la clínica de los pacientes. Esta matriz se caracteriza por presentar los identificadores de cada paciente y los términos extraídos integrados en el corpus global. Además, cada término lleva asociado un número que indica el número de veces que aparece un concepto específicamente en un informe. La captura mostrada a continuación es un ejemplo de los diez primeros pacientes que aparecen en la matriz **(Figura 4.1)**. El número total de pacientes corresponde con el global de informes procesados, donde el número de términos es independiente, estableciendo conteos entre 1 y más de 15 términos en función del informe analizado.

La función *mapeoTerminos* fue desarrollada para la construcción de la matriz, utilizando de nuevo los informes con la etiqueta exonim, con el fin de evitar un tiempo excesivo de computación y generar el análisis en aquellos informes más enriquecidos a nivel de clínica. Esta función integra el uso de la técnica Lookup pero de una forma diferente. En este caso, se utiliza el proceso de búsqueda entre cada informe y el corpus global, de forma que en cada iteración, la información de cada informe se almacena en una lista que contiene datos procesados por el módulo *Counter* de la librería *Collections*. Estos datos están relacionados con el conteo de cada término por informe, los cuales se añaden a la lista junto con el identificador del informe. De esta forma, se obtiene una lista final con todos los identificadores y el conteo relativo a la terminología extraída de los informes. El último paso involucra generar un DataFrame y guardarlo en formato excel para analizar la matriz resultante **(MS-A)**.

Analíticamente, el conteo del concepto puede indicar la importancia del mismo en el contexto clínico del informe. Si observamos la figura 4.1, en los casos de los pacientes p3 y p6, el número de términos es de cuatro. Sin embargo, los pacientes p1 y p2 presentan un recuento mayor en cuanto al término extraído (en este caso de dos). Por ello, deducimos que los resultados obtenidos son muy diversos a nivel de conteo, pudiendo aparecer en la matriz un número de términos de quince o incluso superior en función del caso analizado.

4.6.2. Selección de terminología para su asociación manual a HPO

El mapeo automático de los términos a HPO se realizó mediante la ejecución de un programa informático desarrollado específicamente para esta tarea por el grupo del BSC, mientras que el mapeo manual de los términos lo llevará a cabo un profesional médico con experiencia en

	Id	término 1	término 2	término 3	término 4
p1	16NS1514	miopatia: 2			
p2	16NS1515	neuropatia: 2	polineuropatia: 2		
p3	16NS1601	esteatosis: 1	livedo reticularis: 1	reticularis: 1	vasculitis: 1
p4	17NR0208	retraso: 1	retraso psicomotor: 1		
p5	17NR0209	ataxia: 1			
p6	17NR0216	hernia: 1	hernia discal: 1	charcot: 1	charcot marie: 1
p7	17NR05151	hipoacusia neurosensorial: 1			
p8	17NR05225	discapacidad intelectual: 1			
p9	17NR1053	esclerosis: 1	esclerosis lateral: 1		
p10	17NR1054	discinesia: 1	discinesia ciliar: 1	discinesia ciliar primaria: 1	

Figura 4.1: Captura de la matriz de distribución terminológica. Se observan los 10 primeros pacientes donde se distingue el identificador y el conteo de términos por informe.

este ámbito. Teóricamente, el mapeo automático será un proceso más rápido, pero quizás con una menor fiabilidad a la hora de establecer la asociación de HPO a los términos. Por otra parte, el hecho de realizar un mapeo manual por medio de un profesional especializado, aunque inicialmente implique un proceso más lento y dedicado, en teoría las probabilidades y fiabilidad de los HPO asociados a los términos son más altas en comparación con el mapeo automático.

En este apartado, se explican los procedimientos llevados a cabo para seleccionar los 500 términos que presenten una menor frecuencia de aparición con respecto a la totalidad de conceptos extraídos de los informes. Se seleccionaron aquellos con una frecuencia menor, debido a que son aquellos sobre los cuales los analistas tienen un control menor y, por tanto, su anotación a HPO resulta más complicada en función del contexto del diagnóstico clínico del paciente. Este hecho no ocurre con aquellos términos cuya frecuencia es más alta, puesto que los analistas están más acostumbrados a lidiar con ellos y tienen más facilidades a la hora de decidir que procedimientos seguir cuando se encuentren con casos de esta tipología. Por ello, para generar anotaciones fiables en estos casos, se decidió realizar un mapeo manual de estos términos a HPO.

Este proceso, a diferencia de la construcción de la matriz de distribución descrita en el apartado anterior, toma como referencia el corpus generado por la estrategia lookup sin curar y el corpus global. El diseño del proceso también involucra la técnica lookup pero, al igual que en el apartado anterior, el enfoque de su uso es distinto. En este caso, utilizamos la técnica para ir contando los términos del corpus lookup *sin curar* que aparezcan en el corpus global, de tal modo que el conteo de un término en particular vaya aumentando conforme aparezca en el corpus lookup. Para ejecutar esta estrategia se desarrolló la función *subsetManual*, que funcionalmente permite generar el documento de los 500 términos con menor frecuencia y además, tras ejecutarse, genera dos diccionarios diseñados para establecer un conteo global de términos. Para este paso se utilizó la función *OrderedDict* del módulo *Collections* que permitió realizar el conteo en modo ascendente y descendente. Posteriormente se construyó un DataFrame con estos datos y se guardaron como documentos excel. Estos diccionarios nos permiten analizar cuáles son los conceptos que poseen una mayor y menor frecuencia de aparición respecto de la clínica de los pacientes, siendo posible observar conceptos más generales y términos muy concretos que por lo general están relacionados con enfermedades raras **(Figura 4.2)**.

	término	Frecuencia descendente	término	Frecuencia ascendente
t1	autismo	8072	abdomen distendido	1
t2	cancer	6196	acalasia esofagica	1
t3	epilepsia	3236	acidemias propionicas	1
t4	discapacidad intelectual	2335	adenitis abscesificada	1
t5	retraso psicomotor	1799	adenocarcinoma colon sigmoide	1
t6	diabetes	1328	adenocarcinoma moderadamente diferenciado	1
t7	microcefalia	1218	adenopatias palpables	1
t8	malformaciones	1211	adenosis esclerosante	1
t9	retraso mental	1153	afasia no fluente	1
t10	ataxia	1147	afasia progresiva primaria	1

Figura 4.2: Diccionarios de términos ordenados en modo ascendente y descendente según su frecuencia de aparición en los informes. Los términos generales de enfemedades como autismo, cáncer o epilepsia son más comunes. Los términos más infrecuentes como abdomen distendido o acalasia esofágica son relativos a fenotipos clínicos muy específicos.

Además, en la última parte del código, se utilizó el módulo *FreqDist* de la librería NLTK para localizar aquellos 500 términos menos frecuentes. Este paso se llevó a cabo mediante el uso la función *most_common* de *FreqDist*, la cual genera una lista de tuplas que se utiliza para guardar la clave de la tupla en cada iteración como texto plano **(MS-A)**.

Por otro lado, para representar correctamente el conteo del vocabulario extraído a nivel global y entender mejor su distribución en el espacio, se construyó una imagen de gráficos de barras que permitió observar como se iba modificando la frecuencia de los términos conforme se representaban ventanas de frecuencia cada vez más pequeñas en función del rango de conteo de los términos. Los datos para construir la imagen proceden del diccionario que devuelve la función al ejecutarse, el cual es generado por la función *items* del módulo *FreqDist* de NLTK.

A nivel de código, el diccionario se referencia como *all_items*, pero en el *main_code* se le denominó *all_terms*, simbolizando que este diccionario contiene el conteo global de todos los términos extraídos de los informes. La figura resultante del procesamiento de *all_terms* representa la distribución de todos los conceptos en cuatro categorías terminológicas dependientes de la frecuencia y el rango de conteo establecido, el cual siempre fue de tres grupos distintos que iban variando conforme la frecuencia iba disminuyendo en cada categoría **(Figura 4.3)**.

En la primera categoría se dividió el conteo de los términos en frecuencias mayores de 500, entre 500 y 50 y menor de 50. La segunda categoría se dividió en frecuencias de entre 50 y 25, 25 y 10 y menor de 10. En la tercera categoría el conteo de los términos ya revelaba que las palabras que aparecían 3 o menos veces eran las que se encontraban en mayor proporción. No obstante, la división se realizó entre frecuencias de entre 10 y 5, 5 y 3 y menor de 3. La última ventana estaba dirigida al análisis de la distribución de los términos que se repetían 3, 2 o una única vez, por lo que necesariamente bajo las condiciones anteriormente establecidas, la selección de los 500 términos para el mapeo manual provenía directamente de los conceptos que se encontraban en estos rangos de frecuencia.

Para construir la imagen de distribución terminológica, se utilizó el diccionario *all_terms* que devuelve la función *subsetManual*, el cual integra la lista completa de términos en formato *key-value*, es decir, el concepto relativo al fenotipo del paciente, junto con su frecuencia de aparición global. Los procedimientos seguidos para ir estrechando rangos de frecuencia cada vez menores

están indicados en el *main_ code* e implican la construcción de distintos diccionarios a partir del inicial. De esta forma, la estructura establecida en los rangos de frecuencia implica ir reduciendo paulatinamente el número de veces que aparece un concepto a nivel global.

Finalmente obtenemos la cuarta categoría (D), la cual es representativa de los 500 términos seleccionados con menor frecuencia. Si observamos la figura 4.3 nos damos cuenta de que la proporción de términos que aparecen una única vez ya supera los 500 términos necesarios para el subset manual. Sin embargo, en la práctica se encontraron algunos falsos positivos que necesariamente tuvieron que eliminarse, obligando a incorporar conceptos que aparecían dos o tres veces con el fin de completar la selección del subconjunto de términos.

El último paso consistía en graficar la distribución de estas categorías utilizando como base principal el diagrama de barras, el cual resultó muy útil a la hora de visualizar la frecuencia de los términos en función del número total de conceptos.

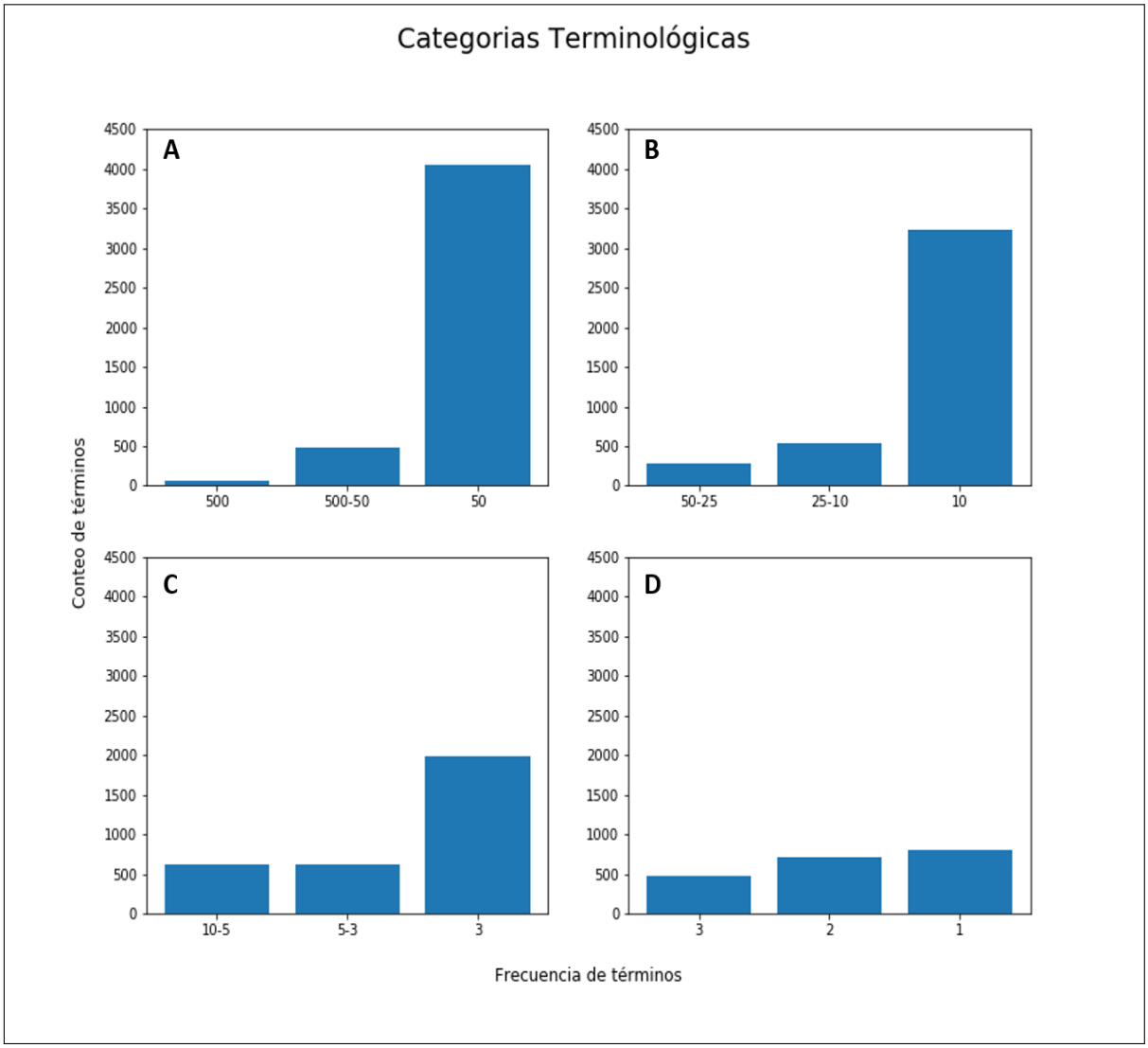

Figura 4.3: Representación de la distribución terminológica en los informes a nivel global. Se observan cuatro categorías diferentes según el rango de frecuencia: **A)** Términos comprendidos entre frecuencias mayores de 500, 500 y 50 y menor de 50. **B)** Términos comprendidos entre frecuencias de 50 y 25, 25 y 10 y menor de 10. **C)** Términos comprendidos entre frecuencias de 10 y 5, 5 y 3 y menor de 3. **D)** Términos con frecuencias de 3, 2 y 1. Categoría representativa de la selección de los 500 términos con menor frecuencia para seleccionar el subconjunto que se asociará manualmente a terminología HPO.

4.7. Exomiser

El principal desafío de WES radica en el descubrimiento de la variante genética causante de la enfermedad. Un exoma individual suele albergar más de 30.000 variantes en comparación con el genoma de referencia, de las cuales se pronostica que aproximadamente 10.000 producirán alteraciones a nivel de secuencia (sustituciones aminoacídicas, cambios de residuos en sitios conservados genéticamente o pequeñas inserciones y deleciones). Incluso después de filtrar variantes comunes, son necesarios métodos adicionales para predecir qué variantes pueden tener consecuencias funcionales graves y priorizarlas para su posterior validación [31] [32].

Exomiser es un algoritmo que primero filtra las variantes según su rareza, ubicación genómica (p. ej. adyacencia a un exón) y la compatibilidad con el modo de herencia. Posteriormente clasifica el resto de genes con variantes identificadas en función del *variant score* (frecuencia y patogenicidad de la variante) y el *phenotype relevant score* , el cual se calcula según la similitud semántica con la enfermedad humana (términos HPO) y las manifestaciones analizadas por el modelo intrínseco de inferencia fenotípica de exomiser [16].

En resumen, exomiser proporciona una manera simple y altamente efectiva de priorizar genes candidatos en función de modelos comparativos que albergan medidas ya existentes, como el grado de patogenicidad de la variante o la información proporcionada por otros atributos como MAF (minor allele frecuency).

4.7.1. Procedimientos previos a la ejecución de exomiser

Los resultados de exomiser se obtuvieron a partir de dos vías distintas: a) ejecutándolo desde el servidor de NIMGenetics donde se encuentra previamente instalado y configurado y b) a través de la plataforma *Health29* que integra exomiser dentro de sus funcionalidades. Su ejecución se realizó sobre los diez casos seleccionados por el departamento médico de la empresa, en los cuales la variante causal del fenotipo del paciente era conocida. La descripción del fenotipo se completó mediante la terminología HPO generada automáticamente y la asociada de forma manual por un profesional médico, la cual se puede visualizar en los anexos del trabajo **(MS-B)**.

En primer lugar, exomiser fue ejecutado desde el servidor de NIMGenetics. Para llevar a cabo este proceso, la función *makeTemplate* fue desarrollada con el objetivo de generar los *templates* necesarios para poder ejecutar exomiser **(MS-A)**. Estos *templates* son documentos en formato *yml* estructurados en dos secciones. En la primera sección se definen las características generales del input y en la segunda se indica la ubicación donde almacenar el output generado por exomiser. La función modifica el *template* original y lo adapta a cada caso de estudio.

En la sección del input es necesario indicar la localización del VCF correspondiente a cada caso de estudio, el modo de análisis y las bases de datos de frecuencias poblacionales seleccionadas para reunir información con respecto a las variantes genéticas, de forma que exomiser únicamente priorizará las variantes filtradas por los requisitos aquí establecidos. Por último, como se ha comentado en el párrafo anterior, en esta sección también se incluyó la información fenotípica relativa al paciente mediante una lista de términos HPO.

En la segunda sección se indica la ruta donde almacenar el output generado por exomiser, la cual es modificada en el documento yml cada vez que la función se ejecuta. El último paso involucra seleccionar el formato en el que guardar los resultados, el cual puede ser tsv-gene, tsv-variant, html o vcf. Para obtener una perspectiva más amplia de los resultados se decidió seleccionar una combinación de formatos que asegurase obtener el máximo de información respecto de la priorización de las variantes candidatas.

4.7.2. Ejecución de exomiser desde servidor

Exomiser fue ejecutado utilizando el archivo jar con el que viene instalado. Para acceder y lanzarlo desde la terminal el comando utilizado fue el siguiente:

java -Xms2g -Xmx4g -jar exomiser-cli-7.2.3.jar –analysis example.yml

En esta situación, para evitar sobrecargar la RAM del ordenador, la opción *PASS_ ONLY* de exomiser fue seleccionada para restringir el análisis únicamente a las variantes que cumplieron los requisitos de filtrado.

4.7.3. Análisis de los resultados de exomiser

En este apartado se evaluarán los resultados obtenidos por exomiser y la plataforma *Health29*. La terminología HPO aportada es imprescindible para representar las anormalidades fenotípicas que presenta el paciente y para que la priorización de variantes establecida por exomiser tenga más probabilidades de resultar exitosa.

Los términos HPO asociados a la clínica del paciente fueron seleccionados manualmente por un profesional médico de la empresa y mapeados automáticamente por el grupo del BSC. En esta situación, exomiser fue ejecutado dos veces por cada caso de estudio, utilizando tanto la terminología asociada de forma manual como los términos obtenidos por mapeo automático. De esta forma, se obtuvieron dos listas de variantes candidatas en función de los datos fenotípicos aportados, sin olvidar los resultados obtenidos por *Health29*. Para mantener la privacidad de los pacientes, los casos fueron configurados mediante el año y el tipo de informe, asociando un número de referencia correspondiente al caso de estudio.

La ejecución de exomiser utilizando la plataforma *Health29* tiene un enfoque más clínico en comparación con su ejecución desde servidor. *Health29* aporta una lista de genes candidatos junto con sus condiciones relacionadas a nivel de síndrome o enfermedad. Por otra parte, la configuración utilizada para ejecutarlo a través de servidor permite obtener una lista de variantes que incluye un conjunto de atributos funcionales por cada una de ellas, lo cual resulta imprescindible a nivel informativo para el profesional médico que solicitó inicialmente el análisis genético.

Para analizar los resultados dos funciones fueron desarrolladas: *filterVariants* y *locVariant*. La primera genera una tabla que muestra las diez primeras variantes priorizadas por exomiser. Esta tabla es el resultado de aplicar dos filtros: el primer filtro omite las variantes catalogadas con un *variant score* de cero, ya que son variantes no patogénicas sin impacto funcional en la proteína, y el segundo elimina los genes de tipología compleja, los cuales presentan características muy diferentes en comparación con genes de tipo mendeliano **(MS-A)**.

La segunda función se diseñó para buscar la variante genética seleccionada por los analistas en la lista de variantes candidatas generada por exomiser. Esta función fue utilizada en los casos en que la variante no aparecía entre las diez primeras posiciones de la lista, con el fin de conocer su posición en la lista global y así tratar de comprender el porqué de ese posicionamiento. Estos resultados se pueden visualizar en el *main_ code* del repositorio github de este trabajo.

Estos resultados junto con los obtenidos por la plataforma *Health29* son suficientes para llevar a cabo dos tareas de forma simultánea: a) analizar si la lista de variantes priorizada por exomiser es significativa respecto de la variante seleccionada por los analistas como causante del fenotipo del paciente y b) testear la funcionalidad de la plataforma *Health29* como herramienta útil en el ámbito del diagnóstico clínico de enfermedades genéticas raras.

Caso de estudio: 16NS-1

Este primer caso es un ejemplo de la importancia que tiene la cantidad de información fenotípica que se aporte a la hora de ejecutar exomiser y *Health29*. La clínica del paciente aportada en este caso era muy pobre, obteniéndose únicamente dos términos HPO asociados a los conceptos médicos extraídos del informe. Inicialmente se sospechó que el paciente padecía de discinesia ciliar primaria, donde el correspondiente análisis de secuenciación de los genes asociados adjudicó una variante en el gen CCDC29, pero ni exomiser ni *Health29* fueron capaces de generar una lista de genes que la incluyera, por lo que se dedujo que sin más información relativa al fenotipo del paciente no sería posible generar una lista de variantes candidatas.

Caso de estudio: 16NS-2

Este caso de estudio estaba relacionado con un paciente que presentaba un deterioro cognitivo frontal progresivo. El proceso global resultó en la asociación de doce términos HPO a la terminología médica extraída del informe de clínica del paciente. Por tanto, inicialmente se confirmaba que la calidad del fenotipo del paciente era óptima.

Los analistas identificaron tres posibles variantes causantes del fenotipo del paciente: una variante *missense* tipo VUS del gen UBQLN2 y dos variantes del gen STUB1. Una de ellas compartía las mismas características que la primera y la otra, que producía un cambio de fase, estaba considerada como probablemente patogénica. Tanto exomiser como *Health29* priorizaron estas variantes entre las diez primeras candidatas. En el caso de exomiser, el *variant score* calculado para las variantes superaba el 0.95, lo cual indicaba su posible relación a nivel patogénico con el fenotipo del paciente. Además, la variante UBQLN2, indicada por los analistas como la que mayor probabilidad tenía de ser la causante de la enfermedad, aparecía en primera posición tanto en las listas de exomiser como en *Health29* **(Figura 4.4)**.

Este caso, a diferencia del anterior, resulta en un suministro adecuado de terminología HPO para describir el fenotipo del paciente. Además, se puede afirmar que la información obtenida tras completar los procedimientos, podría resultar útil para los analistas en el caso de que tuviera que lidiar con algún caso similar, ya que las variantes identificadas fueron todas incluidas en la lista de variantes candidatas priorizada por exomiser.

Caso de estudio: 16NS-3

Este caso corresponde con un ExoNIM Dirigido de un grupo de genes implicados en cardiomiopatía. Las variantes genómicas inicialmente identificadas por los analistas en este estudio fueron dos variantes *missense* en los genes TNN y MYPN de tipo VUS. Por otro lado, la calidad del fenotipo era muy baja, ya que únicamente se aportó un término HPO como en el primer caso, por lo que inicialmente se pensó que los resultados eran incongruentes.

Posteriormente, se demostró mediante validación por Sanger que una variante *missense* del gen MYH7 explicaba el fenotipo patológico del paciente. Este caso, por tanto, era una excepción, puesto que a pesar de que la calidad del fenotipo era mínima, la lista de variantes candidatas priorizada por exomiser albergaba la variante MYH7 en primera posición, tras asociar la terminología HPO manualmente, y en tercera posición, tras generar los HPO automáticamente. Además, aunque ya quedaron descartadas, las variantes identificadas inicialmente también aparecían en las listas. Por otro lado, *Health29* no generó una lista de genes en la cual apareciese la variante MYH7 seleccionada por los analistas **(Figura 4.5)**.

Este caso es interesante y se considera una excepción, debido a que exomiser fue capaz de priorizar una lista de variantes en la cual se incluía la variante causal a pesar de las condiciones

fenotípicas de partida. Este hecho puede explicarse por el tipo de enfermedad que manifiesta el paciente, cuyo fenotipo puede explicarse restringiendo su contexto genético a variantes genéticas muy específicas. No obstante, esta situación es poco común, puesto que por lo general si el fenotipo es pobremente descrito, exomiser no es capaz de priorizar una lista de genes significativa, lo que efectivamente ocurre en este caso con *Health29*.

Caso de estudio: 17NR-1

Este caso corresponde con un estudio genético dirigido a la identificación de mutaciones en genes potencialmente causales del fenotipo del paciente mediante secuenciación masiva del exoma humano completo del paciente y sus progenitores, por lo que el estudio se trataba de un ExoNIM Trío. El paciente había sido diagnosticado con un trastorno del espectro autista y rasgos dismórficos y, aunque los progenitores no presentaban antecedentes de interés en relación con este fenotipo del paciente, este enfoque se llevó a cabo con el fin de dilucidar posibles variantes heredadas de los progenitores que cumplieran un patrón de herencia autosómico recesivo.

La variante identificada en este estudio como causante del fenotipo del paciente fue una variante de *novo* del gen PACS1, la cual se trataba de una variante *missense* registrada como variante patogénica en bases de datos como HGMD y Clinvar asociada a discapacidad intelectual y autismo. En este estudio se asociaron seis términos HPO manualmente y cinco de forma automática, por lo que la calidad del fenotipo se encontraba en un rango óptimo respecto del número de términos HPO recomendados.

Exomiser incluyó la variante PACS1 en tercera posición cuando la asociación de los términos HPO se realizó manualmente. Sin embargo, aunque el mapeo automático asoció un número de términos óptimo en relación con la calidad del fenotipo, algunos de los HPO asociados eran distintos a los incluidos manualmente, lo cual produjo que exomiser priorizara una lista de variantes candidatas que no explicasen correctamente el fenotipo del paciente. Por su parte *Health29* posicionó la variante en primer lugar en la lista de variantes generada, resultando el proceso exitoso junto con la ejecución de exomiser con los términos HPO asociados de forma manual **(Figura 4.6)**.

Por otro lado, como se puede observar en la figura representativa de este caso de estudio, el índice de cada variante candidata en la tabla no siempre comienza en cero, sino que lo hace en número más elevados. Este hecho se explica porque los genes de tipología compleja, a los cuales exomiser ubicaba en primeras posiciones, son filtrados previamente a la construcción de la tabla. Algo que no ocurre en *Health29* debido a que internamente ya omite este conjunto de genes de tipología compleja. Además, se buscó la localización de la variante PACS1 en la lista priorizada por exomiser con los HPO generados automáticamente y se observó que su posición era muy superior a la de las diez primeras variantes. Curiosamente su *variant score* era muy alto, pero su *pheno score* era considerablemente más bajo en comparación con el obtenido en la lista generada en primera instancia con los HPO asociados de forma manual.

Caso de estudio: 17NR-2

Este caso al igual que el anterior correspondía con un estudio de tipo ExoNIM Trío. El paciente manifestaba leucoencefalopatía, mientras que sus progenitores no presentaban antecedentes de interés. En este contexto, las variantes genómicas identificadas por los analistas como posibles causantes del fenotipo del paciente fueron dos variantes *missense* de tipo patogénica registradas en la literatura del gen UBA5.

El número de términos HPO asociados tanto manualmente como mapeados automáticamente fue de siete, por lo que la calidad del fenotipo se encontraba dentro de rangos óptimos. Sin

embargo, en este caso ocurría algo similar a lo acontecido en el caso anterior, donde el mapeo de HPO de forma automática sustituyó algunos de los HPO asociados manualmente por otros que no eran los más idóneos, provocando que la lista priorizada por exomiser no cosechara buenos resultados respecto de la explicación del fenotipo patológico del paciente. Por su parte, la lista priorizada a partir de la asociación manual de términos HPO, posicionó ambas variantes en primera y segunda posición, mientras que en *Health29* la variante también se ubico en segunda posición, pudiendo afirmar que el proceso resultó exitoso, a excepción de los resultados obtenidos mediante mapeo automático **(Figura 4.7)**.

En este estudio, al igual que en el anterior, las dos variantes del gen UBA5 aparecen en posiciones superiores respecto de las diez primeras variantes en la lista priorizada con los HPO generados automáticamente. Se observó que el *variant score* era idéntico al obtenido para las variantes posicionadas en la lista priorizada con los HPO asociados manualmente, pero que su *pheno score* decaía drásticamente debido a que alguno de los HPO adjudicados eran erróneos, lo cual explica que ocupasen dichas posiciones en la lista de variantes candidatas.

Caso de estudio: 17NR-3

Este caso de estudio se trababa de un ExoNIM Dirigido que tenía como objetivo identificar variantes genómicas del gen POGZ. La presencia de mutaciones en este gen se asocian con un patrón de herencia autosómico dominante que produce el síndrome de *White-Sutton*, el cual se sospechaba que manifestaba el paciente. Tras la consecución del estudio, los analistas identificaron una variante patogénica en este gen, la cual fue adjudicada como la variante que explicaba el fenotipo del paciente.

En este caso, la calidad del fenotipo es muy reducida, tanto por la parte de asociación manual como por la de mapeo automático, con tan solo tres términos HPO. Por su parte, la ejecución de exomiser utilizando los HPO asociados manualmente situaba a la variante POGZ en quinta posición presentando un *variant score* de 0.95, mientras que *Health29* la ubicaba en tercer lugar. Por otro lado, la ejecución de exomiser utilizando los HPO generados automáticamente no resultó exitosa, ya que el mapeo adjudicó un HPO erróneo al fenotipo, obteniéndose una lista de genes que no era capaz de explicar el fenotipo patológico que presentaba el paciente **(Figura 4.8)**.

Caso de estudio: 17NR-4

Este caso de estudio se trata de un paciente con probable síndrome de microcefalia-linfedema y displasia de tipo coriorretiniana cuya progenitora mostraba síntomas similares. Al igual que en otros casos previamente comentados, se llevó a cabo un enfoque del tipo ExoNIM Trío.

La variante genética identificada por los analistas fue KIF11, la cual estaba asociada a un patrón de herencia autosómico dominante y explicaba los síntomas anteriores. La variante se encontraba en heterocigosis tanto en el paciente como en su progenitora, por lo que el resto de variantes quedaban descartadas, ya que la variante causante había sido heredada. Esta variante afectaba al sitio donador de *splicing* y estaba considerada como probablemente patogénica. La diferencia encontrada en este caso en comparación con otros previos, es que la variante genómica no se manifestaba como una variante de *novo* sino que estaba sujeta a un patrón de herencia existente entre familiares.

Aunque los términos HPO asociados manualmente y los generados automáticamente no eran completamente idénticos, exomiser priorizó en ambos casos listas de variantes en las cuales aparecían entre las diez primeras candidatas la variante seleccionada por los analistas. Se dedujo por tanto que estos cambios entre HPO no afectaron en el proceso de priorización de la variante. Por su parte *Health29* posicionó a la variante KIF11 en primera posición **(Figura 4.9)**.

Caso de estudio: 18NR-1

Este caso de estudio corresponde con un paciente que presentaba numerosos síntomas clínicos relacionados con el síndrome de *Marden-Walker* sin antecedentes familiares. No obstante, se llevó a cabo una aproximación del tipo ExoNIM Trío con el fin de dilucidar un posible patrón de herencia recesivo.

La variante genómica identificada por los analistas como responsable de provocar el fenotipo del paciente fue una variante *missense* del gen SETBP1, la cual fue detectada en heterocigosis de *novo* y estaba considerada como variante patogénica en las bases de datos HGMD y Clinvar.

La calidad del fenotipo estaba bien cubierta, aportando siete y seis HPO para la asociación de términos manual y automática respectivamente. En este caso, aunque el proceso global resultó exitoso, se observó que existía una pequeña variación entre el *pheno score* adjudicado a la variante SETBP1 de la lista generada mediante los HPO asociados manualmente respecto de la incluida en la priorizada con los HPO mapeados de forma automática **(Figura 4.10)**.

Este hecho permitió deducir que en algunos casos no es del todo necesario describir el fenotipo tan concienzudamente, puesto que se obtienen resultados similares e incluso con una puntuación superior. Sin embargo, como ya se ha comentado en otras ocasiones, siempre es recomendable detallar la descripción del fenotipo del paciente de una forma precisa con los HPO apropiados.

Caso de estudio: 18NR-2

Este caso representa el estudio genético de un paciente que manifestaba encefalopatía coreoatetósica y retraso del desarrollo, cuyos progenitores no presentaban antecedentes de interés respecto del fenotipo del paciente. No obstante, al igual que en el caso anterior, se llevó a cabo una aproximación del tipo ExoNIM Trío para estudiar si existía un posible patrón de herencia entre familiares.

Inicialmente se identificaron mutaciones en los genes TUBB4A y CACNA1B en ambos progenitores, lo cual implicaba considerar estas variantes como posibles candidatas. Sin embargo, el estudio genético no reveló alteraciones de estos genes en el paciente, pero sí en el gen ANO3, lo cual hizo que la posibilidad inicial fuese descartada. La variante del gen ANO3 se trataba de una variante *missense* considerada como probablemente patogénica que se estaba manifestando en heterocigosis de *novo* en el paciente. Además, la variante se asociaba con un patrón de herencia autosómico dominante a la distonía 24, el cual estaba ausente en ambos progenitores.

La calidad del fenotipo se encontraba dentro de rangos óptimos respecto de la terminología HPO aportada, siendo de seis el número de términos tanto para los HPO asociados manualmente como para los generados de forma automática. La única diferencia encontrada en relación con la terminología es que uno de los HPO era distinto, lo cual implicaba que la posición de la variante en el caso de la lista priorizada utilizando los HPO mapeados automáticamente fuera inferior en una posición y que tuviera un *pheno score* menor.

Las variantes que inicialmente sugerían indicios de provocar el fenotipo del paciente se ubican en primeras posiciones en las listas generadas por exomiser y *Health29*, mientras que la variante ANO3, identificada como verdadera causante del fenotipo, aparece siempre entre las diez primeras posiciones tras la quinta posición, lo cual confirma que el proceso fue exitoso en todos los casos y que, a pesar de existir alteraciones genómicas potencialmente heredables en los progenitores, la consecución del análisis confirmaría otros resultados que previamente no se consideraban y que descartaban las sospechas iniciales **(Figura 4.11)**.

Caso de estudio: 18NR-3

Este último caso corresponde con un paciente que manifiesta acondroplasia y retraso psicomotor cuyos progenitores carecen de antecedentes de interés. El protocolo llevado a cabo en este estudio es similar al caso anterior, realizando un enfoque del tipo ExoNIM Trío para estudiar los posibles patrones de herencia existentes entre familiares.

Los analistas identificaron a la variante *missense* del gen FGFR3 en heterocigosis de *novo* como la causante del fenotipo del paciente. Esta variante estaba descrita en bases de datos como HGMD y Clinvar como variante patogénica asociada a acondroplasia.

La calidad del fenotipo en cuanto al número de términos HPO asociados manualmente y mapeados de forma automática se encuentra entre rangos óptimos, siendo de doce y seis términos respectivamente. La diferencia del número de términos radica en la consecución del mapeo automático, el cual no generó el mismo número de HPO que asociándolos manualmente. No obstante, la variante FGFR3 se localiza en primera y segunda posición en ambas listas generadas por exomiser variando ligeramente el *pheno score*, lo que tiene sentido ya que existe una disminución del número de HPO debido al mapeo automático. Por su parte *Health29* posiciona la variante directamente en primer lugar **(Figura 4.12)**.

Comentar que ante una variante genómica que puede presentar mutaciones asocidas a numerosas patologías (acondroplasia, síndrome de *Crouzon*, hipoacondroplasia, síndrome de *LADD*, síndrome de *Muenke* o displasia de *SADDAN*), la información fenotípica aportada en forma de HPO fue imprescindible para obtener resultados exitosos respecto de priorizar una lista de variantes candidatas que puedan explicar el fenotipo del paciente.

A

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chrX	56591569	missense_variant	UBQLN2	0.996715	0.787443	0.996715	0.982886
1	chr14	92537354	frameshift_elongation	ATXN3	0.950000	0.706157	0.950000	0.941401
2	chr16	730628	missense_variant	STUB1	1.000000	0.655719	1.000000	0.937705
4	chr16	732171	frameshift_truncation	STUB1	0.950000	0.655719	1.000000	0.937705
5	chr7	73795169	missense_variant	CLIP2	0.987958	0.657653	0.987958	0.932205
6	chr9	135218103	missense_variant	SETX	0.881138	0.749208	0.881138	0.930327
7	chr6	170871049	disruptive_inframe_insertion	TBP	0.850000	0.764932	0.850000	0.921954
9	chr22	29886606	missense_variant	NEFH	0.997507	0.614849	0.997507	0.905837
10	chr9	140773612	splice_donor_variant	CACNA1B	0.900000	0.605219	1.000000	0.899043
11	chr9	140777306	missense_variant	CACNA1B	1.000000	0.605219	1.000000	0.899043

B

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chrX	56591569	missense_variant	UBQLN2	0.996715	0.844116	0.996715	0.990432
1	chr16	730628	missense_variant	STUB1	1.000000	0.696180	1.000000	0.958199
3	chr16	732171	frameshift_truncation	STUB1	0.950000	0.696180	1.000000	0.958199
4	chr14	92537354	frameshift_elongation	ATXN3	0.950000	0.724320	0.950000	0.950988
5	chr13	52513198	missense_variant	ATP7B	0.984096	0.663606	0.984096	0.933856
6	chr2	129075877	missense_variant	HS6ST1	1.000000	0.641948	1.000000	0.928801
7	chr9	135218103	missense_variant	SETX	0.881138	0.737492	0.881138	0.922008
8	chr1	89613326	missense_variant	GBP7	0.983721	0.633701	0.983721	0.911588
9	chr9	140773612	splice_donor_variant	CACNA1B	0.900000	0.618100	1.000000	0.910563
10	chr9	140777306	missense_variant	CACNA1B	1.000000	0.618100	1.000000	0.910563

C

Genes detectados	Condiciones relacionadas
UBQLN2	Amyotrophic lateral sclerosis 15, with or without frontotemporal dementia Amyotrophic lateral sclerosis
ATXN3	Machado-Joseph disease
CLIP2	Williams syndrome
STUB1	Spinocerebellar ataxia, autosomal recessive 16

Figura 4.4: Resultados obtenidos tras ejecutar exomiser y *Health29* para el caso de estudio 16NS-2. **A)** Lista de variantes candidatas priorizada por exomiser tras asociar términos HPO manualmente. **B)** Lista de variantes candidatas priorizada por exomiser tras mapeo automático de términos HPO. **C)** Lista de variantes candidatas priorizada por *Health29*.

A

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
1	chr14	23899016	missense_variant	MYH7	1.000000	1.000000	1.000000	0.998149
2	chr10	69881820	missense_variant	MYPN	0.999596	1.000000	0.999596	0.998142
4	chr2	179600304	missense_variant	TTN	0.992592	1.000000	0.992592	0.998019
6	chr2	179640548	missense_variant	TTN	0.987838	1.000000	0.992592	0.998019
7	chrX	31747837	missense_variant	DMD	0.922496	1.000000	0.922496	0.996236
9	chr5	137222922	missense_variant	MYOT	0.997149	0.829720	0.997149	0.988948
10	chr4	996513	splice_region_variant	IDUA	0.900000	0.766280	0.900000	0.949903
11	chr6	110064448	missense_variant	FIG4	0.999797	0.671483	0.999797	0.946524
15	chr6	31238930	missense_variant	HLA-B	0.758000	0.641241	0.997000	0.926458
16	chr6	31239006	missense_variant	HLA-B	0.997000	0.641241	0.997000	0.926458

B

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr10	69881820	missense_variant	MYPN	0.999596	1.000000	0.999596	0.998142
2	chr5	137222922	missense_variant	MYOT	0.997149	1.000000	0.997149	0.998100
4	chr14	23899016	missense_variant	MYH7	1.000000	0.968246	1.000000	0.997427
5	chr2	179600304	missense_variant	TTN	0.992592	0.968246	0.992592	0.997246
7	chr2	179640548	missense_variant	TTN	0.987838	0.968246	0.992592	0.997246
8	chrX	31747837	missense_variant	DMD	0.922496	1.000000	0.922496	0.996236
10	chr4	996513	splice_region_variant	IDUA	0.900000	0.923365	0.900000	0.989801
11	chr6	110064448	missense_variant	FIG4	0.999797	0.807486	0.999797	0.986443
15	chr6	31238930	missense_variant	HLA-B	0.758000	0.770104	0.997000	0.979627
16	chr6	31239006	missense_variant	HLA-B	0.997000	0.770104	0.997000	0.979627

C

Genes detectados	Condiciones relacionadas
FAM20C	Raine syndrome Lethal osteosclerotic bone dysplasia
PIK3R2	Megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome 1 Megalencephaly-polymicrogyria-postaxial polydactyly-hydrocephalus syndrome
CLIP2	Williams syndrome
IMPDH1	Retinitis pigmentosa 10 Leber congenital amaurosis 11 Leber congenital amaurosis Retinitis pigmentosa

Figura 4.5: Resultados obtenidos tras ejecutar exomiser y *Health29* para el caso de estudio 16NS-3. **A)** Lista de variantes candidatas priorizada por exomiser tras asociar términos HPO manualmente. **B)** Lista de variantes candidatas priorizada por exomiser tras mapeo automático de términos HPO. **C)** Lista de variantes candidatas priorizada por *Health29*.

A

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
162	chr7	73477660	missense_variant	ELN	0.994485	0.735234	0.994485	0.970328
163	chr22	24109722	missense_variant	CHCHD10	0.912603	0.757119	0.912603	0.950871
164	chr11	65978677	missense_variant	PACS1	0.998427	0.680921	0.998427	0.950691
166	chr16	49672716	missense_variant	ZNF423	0.998423	0.651352	0.998423	0.934115
167	chr18	48581231	missense_variant	SMAD4	0.997270	0.643719	0.997270	0.928359
168	chr22	51121852	splice_region_variant	SHANK3	0.811295	0.788879	0.811295	0.913936
169	chr12	112036767	disruptive_inframe_insertion	ATXN2	0.845490	0.757119	0.845490	0.912661
170	chr14	73659482	missense_variant	PSEN1	1.000000	0.615655	1.000000	0.908472
171	chr6	32026164	missense_variant	TNXB	0.992000	0.622534	0.992000	0.908308
173	chr1	103496754	missense_variant	COL11A1	0.991646	0.622644	0.991646	0.908132

B

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr7	73477660	missense_variant	ELN	0.994485	0.796732	0.994485	0.984120
163	chr22	24109722	missense_variant	CHCHD10	0.912603	0.793365	0.912603	0.965766
164	chr18	48581231	missense_variant	SMAD4	0.997270	0.690313	0.997270	0.954613
165	chr6	32026164	missense_variant	TNXB	0.992000	0.667424	0.992000	0.940462
166	chr2	129026227	missense_variant	HS6ST1	1.000000	0.660041	1.000000	0.940279
167	chr2	129075877	missense_variant	HS6ST1	1.000000	0.660041	1.000000	0.940279
168	chr20	5294612	missense_variant	PROKR2	0.998809	0.660041	0.998809	0.939661
169	chr12	112036767	disruptive_inframe_insertion	ATXN2	0.845490	0.793365	0.845490	0.938389
170	chr14	73659482	missense_variant	PSEN1	1.000000	0.656838	1.000000	0.938381
171	chr21	45750578	disruptive_inframe_insertion	C21orf2	0.834885	0.793365	0.834885	0.932512

C

Genes detectados	Condiciones relacionadas
PACS1	Schuurs-Hoeijmakers syndrome
NOTCH2	Hajdu-Cheney syndrome
	Alagille syndrome 2
	Acroosteolysis dominant type
ELN	Cutis laxa, autosomal dominant
	Supravalvar aortic stenosis
	Supravalvular aortic stenosis
	Autosomal dominant cutis laxa
	Williams syndrome
	Familial thoracic aortic aneurysm and aortic dissection

Figura 4.6: Resultados obtenidos tras ejecutar exomiser y *Health29* para el caso de estudio 17NR-1. **A)** Lista de variantes candidatas priorizada por exomiser tras asociar términos HPO manualmente. **B)** Lista de variantes candidatas priorizada por exomiser tras mapeo automático de términos HPO. **C)** Lista de variantes candidatas priorizada por *Health29*.

A

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr3	132384868	missense_variant	UBA5	1.000000	0.656829	1.000000	0.938376
1	chr3	132394747	missense_variant	UBA5	0.888318	0.656829	1.000000	0.938376
2	chr12	7361645	missense_variant	PEX5	0.983341	0.653448	0.983341	0.926557
3	chr5	60368955	missense_variant	NDUFAF2	0.994657	0.634968	0.994657	0.920326
4	chr19	39911625	missense_variant	PLEKHG2	0.992780	0.631986	0.992780	0.916717
5	chr4	3076650	frameshift_truncation	HTT	0.950000	0.633110	0.950000	0.882608
6	chr4	3076653	frameshift_variant	HTT	0.950000	0.633110	0.950000	0.882608
7	chr4	3076659	missense_variant	HTT	0.785737	0.633110	0.950000	0.882608
8	chr4	3076662	missense_variant	HTT	0.759000	0.633110	0.950000	0.882608
9	chr4	3076665	missense_variant	HTT	0.644328	0.633110	0.950000	0.882608

B

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
1	chr21	47422230	frameshift_variant	COL6A1	0.950000	0.785486	0.950000	0.973435
2	chr5	60368955	missense_variant	NDUFAF2	0.994657	0.741998	0.994657	0.972329
3	chr19	39911625	missense_variant	PLEKHG2	0.992780	0.737405	0.992780	0.970526
5	chr4	103808519	missense_variant	CISD2	1.000000	0.651608	1.000000	0.935162
7	chr21	34923650	missense_variant	SON	1.000000	0.636529	1.000000	0.924985
8	chr12	7361645	missense_variant	PEX5	0.983341	0.644344	0.983341	0.919851
9	chr12	24048965	splice_region_variant	SOX5	0.900000	0.699956	0.900000	0.904909
10	chr6	29911296	missense_variant	HLA-A	1.000000	0.575765	1.000000	0.867666
11	chr6	29911319	splice_region_variant	HLA-A	0.900000	0.575765	1.000000	0.867666
16	chr4	3076650	frameshift_truncation	HTT	0.950000	0.615663	0.950000	0.862475

C

Genes detectados	Condiciones relacionadas
HTT	Huntington disease Juvenile Huntington disease Huntington disease
UBA5	Epileptic encephalopathy, early infantile, 44 ?Spinocerebellar ataxia, autosomal recessive 24
ARSD	Ninguna enfermedad conocida
GART	Ninguna enfermedad conocida

Figura 4.7: Resultados obtenidos tras ejecutar exomiser y *Health29* para el caso de estudio 17NR-2. **A)** Lista de variantes candidatas priorizada por exomiser tras asociar términos HPO manualmente. **B)** Lista de variantes candidatas priorizada por exomiser tras mapeo automático de términos HPO. **C)** Lista de variantes candidatas priorizada por *Health29*.

A

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
34	chr3	9517375	missense_variant	SETD5	0.931024	0.904502	0.931024	0.990660
224	chr7	73482987	missense_variant	ELN	0.902591	0.895430	0.902591	0.986727
225	chr6	170871051	frameshift_truncation	TBP	0.950000	0.819662	0.950000	0.981229
226	chr6	170871054	frameshift_variant	TBP	0.950000	0.819662	0.950000	0.981229
228	chr1	151400625	stop_gained	POGZ	0.950000	0.817993	0.950000	0.980906
229	chr16	2112607	splice_region_variant	TSC2	0.900000	0.787658	0.900000	0.959481
230	chr9	140773612	splice_donor_variant	CACNA1B	0.900000	0.650247	1.000000	0.934299
231	chr9	140777306	missense_variant	CACNA1B	1.000000	0.650247	1.000000	0.934299
233	chr4	23816235	splice_region_variant	PPARGC1A	0.900000	0.728623	0.900000	0.927639
234	chr1	110256115	missense_variant	GSTM5	0.994000	0.641622	0.994000	0.924835

B

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr3	9517375	missense_variant	SETD5	0.931024	0.774002	0.931024	0.964688
2	chr7	73482987	missense_variant	ELN	0.902591	0.756475	0.902591	0.946053
3	chr16	2112607	splice_region_variant	TSC2	0.900000	0.753123	0.900000	0.942980
4	chr1	110256115	missense_variant	GSTM5	0.994000	0.651475	0.994000	0.931652
6	chr8	1626783	missense_variant	DLGAP2	0.999998	0.629312	0.999998	0.919609
41	chr6	31238859	missense_variant	HLA-B	0.957551	0.604437	0.997000	0.895758
42	chr6	31238866	missense_variant	HLA-B	0.955038	0.604437	0.997000	0.895758
43	chr6	31238930	missense_variant	HLA-B	0.758000	0.604437	0.997000	0.895758
44	chr6	31239006	missense_variant	HLA-B	0.997000	0.604437	0.997000	0.895758
45	chr6	31323115	missense_variant	HLA-B	0.723000	0.604437	0.997000	0.895758

C

Genes detectados	Condiciones relacionadas
HYDIN	Ciliary dyskinesia, primary, 5 Primary ciliary dyskinesia
TBP	Parkinson disease, susceptibility to Spinocerebellar ataxia 17 Spinocerebellar ataxia type 17
POGZ	White-Sutton syndrome

Figura 4.8: Resultados obtenidos tras ejecutar exomiser y *Health29* para el caso de estudio 17NR-3. **A)** Lista de variantes candidatas priorizada por exomiser tras asociar términos HPO manualmente. **B)** Lista de variantes candidatas priorizada por exomiser tras mapeo automático de términos HPO. **C)** Lista de variantes candidatas priorizada por *Health29*.

A

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr10	97402858	missense_variant	ALDH18A1	0.984244	0.645947	0.984244	0.921672
169	chr10	94410275	splice_donor_variant	KIF11	0.900000	0.704208	0.900000	0.908645
174	chr6	31239006	missense_variant	HLA-B	0.997000	0.577401	0.997000	0.866450
177	chr6	31323202	missense_variant	HLA-B	0.001000	0.577401	0.997000	0.866450
178	chr6	31323962	frameshift_elongation	HLA-B	0.950000	0.577401	0.997000	0.866450
179	chr6	31324100	missense_variant	HLA-B	0.488000	0.577401	0.997000	0.866450
181	chr1	120510194	missense_variant	NOTCH2	0.835154	0.625004	0.942000	0.865254
183	chr1	120572547	missense_variant	NOTCH2	0.871000	0.625004	0.942000	0.865254
184	chr1	120572572	missense_variant	NOTCH2	0.942000	0.625004	0.942000	0.865254
193	chr1	120611964	missense_variant	NOTCH2	0.815000	0.625004	0.942000	0.865254

B

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr7	92132483	frameshift_elongation	PEX1	0.933753	0.683736	0.933753	0.916398
3	chr4	1808659	missense_variant	FGFR3	0.985399	0.633153	0.985399	0.912368
4	chr10	94410275	splice_donor_variant	KIF11	0.900000	0.684387	0.900000	0.890042
5	chr10	97402858	missense_variant	ALDH18A1	0.984244	0.603234	0.984244	0.883016
7	chr2	170083020	missense_variant	LRP2	1.000000	0.586161	1.000000	0.879588
10	chr3	10088266	splice_region_variant	FANCD2	0.900000	0.664021	0.900000	0.867551
13	chr3	10088343	missense_variant	FANCD2	0.686000	0.664021	0.900000	0.867551
15	chr3	10088409	splice_region_variant	FANCD2	0.900000	0.664021	0.900000	0.867551
21	chr3	10089738	splice_region_variant	FANCD2	0.900000	0.664021	0.900000	0.867551
34	chr3	183885694	missense_variant	DVL3	0.997372	0.559865	0.997372	0.844365

C

Genes detectados	Condiciones relacionadas
KIF11	Microcephaly with or without chorioretinopathy, lymphedema, or mental retardation Microcephaly-lymphedema-chorioretinopathy syndrome
HYDIN	Ciliary dyskinesia, primary, 5 Primary ciliary dyskinesia
HLA-A	Birdshot chorioretinopathy
HLA-B	Spondyloarthropathy, susceptibility to, 1 Behçet disease Takayasu arteritis Stevens-Johnson syndrome

Figura 4.9: Resultados obtenidos tras ejecutar exomiser y *Health29* para el caso de estudio 17NR-4. **A)** Lista de variantes candidatas priorizada por exomiser tras asociar términos HPO manualmente. **B)** Lista de variantes candidatas priorizada por exomiser tras mapeo automático de términos HPO. **C)** Lista de variantes candidatas priorizada por *Health29*.

A

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr18	42531907	missense_variant	SETBP1	1.000000	0.673894	1.000000	0.947870
1	chr1	8420514	missense_variant	RERE	1.000000	0.622745	1.000000	0.914418
5	chr1	6111709	missense_variant	KCNAB2	0.981009	0.622745	0.981009	0.899745
7	chrX	132888207	splice_region_variant	GPC3	0.900000	0.693116	0.900000	0.898613
9	chr2	152420160	missense_variant	NEB	0.717769	0.626933	0.959208	0.884700
11	chr2	152584282	missense_variant	NEB	0.959208	0.626933	0.959208	0.884700
12	chr6	3225770	missense_variant	TUBB2B	1.000000	0.589064	1.000000	0.882748
13	chr22	21153431	missense_variant	PI4KA	0.982848	0.593527	0.982848	0.870747
14	chr10	101487264	missense_variant	COX15	0.980254	0.595636	0.980254	0.870533
16	chr6	142691950	frameshift_truncation	ADGRG6	0.950000	0.619586	0.950000	0.867242

B

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr18	42531907	missense_variant	SETBP1	1.000000	0.801660	1.000000	0.985636
1	chr22	21153431	missense_variant	PI4KA	0.982848	0.749710	0.982848	0.971558
2	chr1	8420514	missense_variant	RERE	1.000000	0.732285	1.000000	0.970898
4	chr6	3225770	missense_variant	TUBB2B	1.000000	0.726136	1.000000	0.969037
7	chr1	6111709	missense_variant	KCNAB2	0.981009	0.732285	0.981009	0.965544
9	chrX	132888207	splice_region_variant	GPC3	0.900000	0.793909	0.900000	0.961933
11	chr2	152420160	missense_variant	NEB	0.717769	0.704881	0.959208	0.945213
13	chr2	152584282	missense_variant	NEB	0.959208	0.704881	0.959208	0.945213
14	chr2	167055352	missense_variant	SCN9A	0.966000	0.691928	0.966000	0.941348
16	chr11	792058	missense_variant	SLC25A22	0.979050	0.674888	0.979050	0.938104

C

Genes detectados	Condiciones relacionadas
SETBP1	Schinzel-Giedion midface retraction syndrome
	Mental retardation, autosomal dominant 29
TUBB2B	Cortical dysplasia, complex, with other brain malformations 7
	Dysequilibrium syndrome
MST1	Primary sclerosing cholangitis

Figura 4.10: Resultados obtenidos tras ejecutar exomiser y *Health29* para el caso de estudio 18NR-1. **A)** Lista de variantes candidatas priorizada por exomiser tras asociar términos HPO manualmente. **B)** Lista de variantes candidatas priorizada por exomiser tras mapeo automático de términos HPO. **C)** Lista de variantes candidatas priorizada por *Health29*.

A

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr19	6495816	missense_variant	TUBB4A	0.999295	0.765441	0.999295	0.979074
1	chr9	140777306	missense_variant	CACNA1B	1.000000	0.737432	1.000000	0.972373
3	chr9	141016163	missense_variant	CACNA1B	0.994195	0.737432	1.000000	0.972373
4	chr22	24167605	splice_region_variant	SMARCB1	0.900000	0.817198	0.900000	0.969871
5	chr6	170871051	frameshift_truncation	TBP	0.950000	0.713726	0.950000	0.945593
6	chr6	170871054	frameshift_truncation	TBP	0.950000	0.713726	0.950000	0.945593
8	chr11	26621221	missense_variant	ANO3	0.999983	0.656656	0.999983	0.938263
10	chr19	50364706	missense_variant	PNKP	0.998984	0.655126	0.998984	0.936794
11	chr6	33419613	frameshift_truncation	SYNGAP1	0.950000	0.692028	0.950000	0.932754
12	chr6	33419615	frameshift_elongation	SYNGAP1	0.950000	0.692028	0.950000	0.932754

B

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr22	24167605	splice_region_variant	SMARCB1	0.900000	0.781955	0.900000	0.957112
1	chr19	6495816	missense_variant	TUBB4A	0.999295	0.689336	0.999295	0.954977
2	chr9	140777306	missense_variant	CACNA1B	1.000000	0.649400	1.000000	0.933756
4	chr9	141016163	missense_variant	CACNA1B	0.994195	0.649400	1.000000	0.933756
5	chr6	170871051	frameshift_truncation	TBP	0.950000	0.634009	0.950000	0.883572
6	chr6	170871054	frameshift_truncation	TBP	0.950000	0.634009	0.950000	0.883572
7	chr19	50364706	missense_variant	PNKP	0.998984	0.588640	0.998984	0.881319
9	chr11	26621221	missense_variant	ANO3	0.999983	0.581794	0.999983	0.874680
11	chr6	33419613	frameshift_truncation	SYNGAP1	0.950000	0.618809	0.950000	0.866309
12	chr6	33419615	frameshift_elongation	SYNGAP1	0.950000	0.618809	0.950000	0.866309

C

Genes detectados	Condiciones relacionadas
TUBB4A	Dystonia 4, torsion, autosomal dominant Leukodystrophy, hypomyelinating, 6 Primary dystonia, DYT4 type
CACNA1B	?Dystonia 23
TBP	Parkinson disease, susceptibility to Spinocerebellar ataxia 17 Spinocerebellar ataxia type 17
SYNGAP1	Mental retardation, autosomal dominant 5
CD7	Ninguna enfermedad conocida
ALMS1	Alstrom syndrome Alström syndrome
MCTP2	Distal monosomy 15q
SMC1A	Cornelia de Lange syndrome 2 Cornelia de Lange syndrome
ANO3	Dystonia 24

Figura 4.11: Resultados obtenidos tras ejecutar exomiser y *Health29* para el caso de estudio 18NR-2. **A)** Lista de variantes candidatas priorizada por exomiser tras asociar términos HPO manualmente. **B)** Lista de variantes candidatas priorizada por exomiser tras mapeo automático de términos HPO. **C)** Lista de variantes candidatas priorizada por *Health29*.

A

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
1	chr4	1806119	missense_variant	FGFR3	1.000000	1.0	1.000000	0.998149
3	chr7	27140845	missense_variant	HOXA2	0.999999	1.0	0.999999	0.998149
185	chr1	235894211	missense_variant	LYST	0.997261	1.0	0.997261	0.998102
187	chr6	31239006	missense_variant	HLA-B	0.997000	1.0	0.997000	0.998097
188	chr6	31239111	frameshift_elongation	HLA-B	0.950000	1.0	0.997000	0.998097
189	chr6	31239113	frameshift_truncation	HLA-B	0.950000	1.0	0.997000	0.998097
190	chr6	31239133	splice_region_variant	HLA-B	0.776014	1.0	0.997000	0.998097
191	chr6	31323115	missense_variant	HLA-B	0.723000	1.0	0.997000	0.998097
193	chr6	31324100	missense_variant	HLA-B	0.488000	1.0	0.997000	0.998097
195	chr4	151829506	missense_variant	LRBA	0.981000	1.0	0.981000	0.997797

B

	chr	pos	functional_class	gene	variant_score	pheno_score	gene_variant_score	gene_combined_score
0	chr12	49434351	missense_variant	KMT2D	0.977933	0.924484	0.977933	0.995046
2	chr4	1806119	missense_variant	FGFR3	1.000000	0.847534	1.000000	0.991035
4	chr4	151829506	missense_variant	LRBA	0.981000	0.834693	0.981000	0.987841
5	chr17	17697101	frameshift_truncation	RAI1	0.950000	0.854916	0.950000	0.986912
189	chr10	75519551	splice_region_variant	SEC24C	0.890018	0.844099	0.890018	0.974904
190	chr22	19960512	missense_variant	ARVCF	0.872484	0.844099	0.872484	0.970648
191	chr1	22149986	splice_region_variant	HSPG2	0.857924	0.849248	0.857924	0.968274
194	chr19	15281611	missense_variant	NOTCH3	1.000000	0.703626	1.000000	0.961191
198	chr15	64027048	missense_variant	HERC1	1.000000	0.687365	1.000000	0.954371
199	chr7	286468	stop_gained	FAM20C	0.950000	0.642723	1.000000	0.929332

C

Genes detectados	Condiciones relacionadas
FGFR3	Achondroplasia
	Bladder cancer, somatic
	Colorectal cancer, somatic
	Hypochondroplasia
	LADD syndrome
	Nevus, epidermal, somatic
	Thanatophoric dysplasia, type I
	Thanatophoric dysplasia, type II
	Spermatocytic seminoma, somatic
	Muenke syndrome
	Cervical cancer, somatic
	CATSHL syndrome
	Crouzon syndrome with acanthosis nigricans
	SADDAN

Figura 4.12: Resultados obtenidos tras ejecutar exomiser y *Health29* para el caso de estudio 18NR-3. **A)** Lista de variantes candidatas priorizada por exomiser tras asociar términos HPO manualmente. **B)** Lista de variantes candidatas priorizada por exomiser tras mapeo automático de términos HPO. **C)** Lista de variantes candidatas priorizada por *Health29*.

5

Conclusiones y Trabajo Futuro

Las conclusiones que pueden extraerse de los resultados obtenidos en el presente trabajo son las citadas a continuación:

- El pipeline de preprocesamiento desarrollado permitió extraer el contenido de los informes de clínica en un formato adaptable que posibilitara la ejecución del resto de procedimientos diseñados para este proyecto.
- La extracción de terminología médica de los informes mediante cutext y la estrategia lookup resultó satisfactoria, obteniéndose un set de términos muy diverso.
- El mapeo automático de los conceptos médicos procedentes de los informes de clínica permitió ejecutar exomiser utilizando como parámetro de entrada un fenotipo generado automáticamente.
- La comparativa establecida utilizando los términos HPO asociados manualmente y los obtenidos mediante mapeo automático permitió evaluar la viabilidad de exomiser aplicando descripciones fenotípicas semejantes. Los resultados obtenidos tras el análisis confirmaron el excelente rendimiento del proceso al priorizar la variante causal del fenotipo entre las diez primeras posiciones en la mayoría de los casos.

El trabajo futuro dentro del contexto científico de este proyecto, radica en aumentar el análisis del número de casos de estudio para generar datos estadísticos suficientes como para confirmar o establecer conclusiones distintas a las aportadas en este trabajo. Para ello, los esfuerzos a realizar deberán estar dirigidos a seleccionar casos más arbitrariamente que sirvan para comprobar la funcionalidad del procedimiento desarrollado y considerar si fuera oportuno su aplicación dentro de la dinámica de trabajo de la empresa.

Glosario de Acrónimos

- **NGS**: Next Generation Sequencing
- **WES**: Whole Exome Sequencing
- **HPO**: Human Phenotype Ontology
- **VUS**: Variante de significado incierto
- **NIH**: National Institutes of Health
- **Gestlab**: Base de datos de NIMGenetics
- **BSC**: Barcelona Supercomputing Center
- **Plan TL**: Plan de impulso de las tecnologías del lenguaje
- **NLP**: Procesamiento del lenguaje natural
- **MS-A**: Material Suplementario A
- **MS-B**: Material Suplementario B

Bibliografía

[1] Yang Y et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med*, 2013 Oct 17.

[2] Schieppati A et al. Why rare diseases are an important medical and social issue. *Lancet*, 2008 Jun 14.

[3] Bamshad MJ et al. Exome sequencing as a tool for mendelian disease gene discovery. *Nat Rev Genet*, 2011 Sep 27.

[4] Bao R et al. Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform*, 2014 Sep 21.

[5] Petr Danecek and Adam Auton. The variant call format and vcftools. *Bioinformatics*, 2011 Aug 1.

[6] Botstein D et al. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, 2003 Mar.

[7] Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 2014 Oct 15.

[8] Wen Chen et al. Secondary findings in 421 whole exome-sequenced chinese children. *Hum Genomics*, 2018.

[9] Biesecker LG et al. Diagnostic clinical genome and exome sequencing. *N Engl J Med*, 2014 Jun 19.

[10] McLaren W et al. The ensembl variant effect predictor. *Genome Biol*, 2016 Jun 6.

[11] Trujillano D et al. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur J Hum Genet*, 2017 Feb.

[12] Hoffman-Andrews L et al. The known unknown: the challenges of genetic variants of uncertain significance in clinical practice. *J Law Biosci*, 2017 Jan 22.

[13] Shashi V et al. Practical considerations in the clinical application of whole-exome sequencing. *Clin Genet*, 2016 Feb.

[14] Robinson PN et al. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*, 2008 Nov.

[15] Köhler S et al. The human phenotype ontology in 2017. *Nucleid Acids Res*, 2017 Jan 4.

[16] Robinson PN et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*, 2014 Feb.

[17] Zemojtel T et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med*, 2014 Sep 3.

[18] Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat*, 2012 May.

[19] Köhler S et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*, 2009 Oct.

[20] Groza T et al. The human phenotype ontology: Semantic unification of common and rare disease. *Am J Hum Genet*, 2015 Jul 2.

[21] Köhler S et al. Expansion of the human phenotype ontology (hpo) knowledge base and resources. *Nucleic Acids Res*, 2019 Jan 8.

[22] Jéru I et al. Mutations in nalp12 cause hereditary periodic fever syndromes. *Proc Natl Acad Sci U S A*, 2016 Aug.

[23] McMurry JA et al. Navigating the phenotype frontier: The monarch initiative. *Genetics*, 2016 Aug.

[24] Mungall CJ and McMurry JA. The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*, 2017 Jan.

[25] Frey LJ et al. Ehr big data deep phenotyping. contribution of the imia genomic medicine working group. *Yearb Med Inform*, 2014 Aug 15.

[26] Dhombres F and Bodenreider O. Interoperability between phenotypes in research and healthcare terminologies–investigating partial mappings between hpo and snomed ct. *J Biomed Semantics*, 2016 Feb 9.

[27] Rainer W and Bodenreider O. Coverage of phenotypes in standard terminologies. *Proceedings of the Joint BioOntologies*, 2014.

[28] Bodenreider O. The unified medical language system (umls): integrating biomedical terminology. *Nucleid Acids Res*, 2004 Jan 1.

[29] Köhler S et al. Improving ontologies by automatic reasoning and evaluation of logical definitions. bmc. *Bioinformatics*, 2011 Oct 27.

[30] Köhler S et al. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. *Version 2 F1000Res*, 2013 Feb 1.

[31] Pelak K et al. The characterization of twenty sequenced human genomes. *PLoS Genet*, 2010 Sep 9.

[32] Li MX et al. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet*, 2013.

Material Suplementario A

Figura 5.1: Diagrama de flujo que representa las tareas realizadas según en qué etapa del trabajo. Se pueden observar todos los programas desarrollados para la consecución de las mismas y los directorios donde se almacenan los resultados obtenidos tras su ejecución. La carpeta *Resultados* incluye los directorios generados tras procesar los informes de Gestlab, elaborar el corpus, extraer los conceptos médicos de los textos y utilizar la terminología HPO asociada al fenotipo del paciente junto con la información genética para ejecutar exomiser y la plataforma *Health29*.

Material Suplementario B

Terminología HPO		
Caso de Estudio	**Asociación Manual**	**Mapeo Automático**
16NS-1	HP:0002715, HP:0002086	HP:0002715, HP:0002086
16NS-2	HP:0002145, HP:0001251, HP:0002072 HP:0001260, HP:0002071, HP:0001272 HP:0002145, HP:0002527, HP:0002463 HP:0002015, HP:0000666, HP:0002127	HP:0002145, HP:0001251 HP:0001260, HP:0002015 HP:0002465, HP:0001272 HP:0000666
16NR-3	HP:0001644	HP:0001638
17NR-1	HP:0000729, HP:0001999 HP:0000286, HP:0000414 HP:0001611, HP:0000712	HP:0000729, HP:0000286 HP:0000366, HP:0001611 HP:0000712
17NR-2	HP:0002352, HP:0003429, HP:0002415 HP:0001252, HP:0012736, HP:0000577 HP:0030455, HP:0000338, HP:0000817 HP:0001762, HP:0002808, HP:0003376	HP:0002352, HP:0002415 HP:0001252, HP:0001263 HP:0000577, HP:0001123 HP:0008081, HP:0002808
17NR-3	HP:0001263, HP:0007018 HP:0006919	HP:0001263, HP:0007018 HP:0100710
17NR-4	HP:0000118, HP:0000252, HP:0000532 HP:0001004, HP:0001363, HP:0010537 HP:0000233, HP:0000343, HP:0000581 HP:0025349, HP:0000488, HP:0001622	HP:0000252, HP:0001004 HP:0005486, HP:0001363 HP:0000058, HP:0012745 HP:0001145, HP:0001622
18NR-1	HP:0000158, HP:0011410, HP:0002804 HP:0001298, HP:0001250, HP:0001274 HP:0001166, HP:0000119, HP:0001197	HP:0000158, HP:0002804 HP:0001298, HP:0001250 HP:0001274, HP:0001166
18NR-2	HP:0004322, HP:0100021 HP:0001332, HP:0001263 HP:0001510, HP:0001250	HP:0004322, HP:0100021 HP:0001332, HP:0001263 HP:0001548, HP:0001250
18NR-3	HP:0000924, HP:0000152, HP:0001507 HP:0040064, HP:0000707, HP:0000598 HP:0001574, HP:0001626, HP:0002715 HP:0002086, HP:0003011, HP:0001263 HP:0000403	HP:0001263, HP:0000403 HP:0001548, HP:0000707 HP:0002715, HP:0002086

Tabla 5.1: Lista de términos HPO asociados manualmente por un facultativo médico y mapeados de forma automática por el grupo de colaboración del BSC en función del caso de estudio analizado.